

Logarithmic Regret Bounds for Bandits with Knapsacks

Arthur Flajolet

FLAJOLET@MIT.EDU

Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139

Patrick Jaillet

JAILLET@MIT.EDU

Department of Electrical Engineering and Computer Science, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139

Abstract

Optimal regret bounds for Multi-Armed Bandit problems are now well documented. They can be classified into two categories based on the growth rate with respect to the time horizon T : (i) small, distribution-dependent, bounds of order of magnitude $\ln(T)$ and (ii) robust, distribution-free, bounds of order of magnitude \sqrt{T} . The Bandits with Knapsacks model, an extension to the framework allowing to model resource consumption, lacks this clear-cut distinction. While several algorithms have been shown to achieve asymptotically optimal distribution-free bounds on regret, there has been little progress toward the development of small distribution-dependent regret bounds. We partially bridge the gap by designing a general-purpose algorithm with distribution-dependent regret bounds that are optimal in several important cases that cover many practical applications, including dynamic pricing with limited supply, online bid optimization for sponsored search auctions, and dynamic procurement.

Keywords: Multi-Armed Bandits; Knapsack Constraints

1. Introduction

1.1. Motivation

Multi-Armed Bandit (MAB) is a benchmark model for repeated decision making in stochastic environments with very limited feedback on the outcomes of alternatives. In these circumstances, a decision maker must strive to find an overall optimal sequence of decisions while making as few suboptimal ones as possible when exploring the decision space in order to generate as much revenue as possible, a trade-off coined *exploration-exploitation*. The original problem, first formulated in its predominant version in Robbins (1952), has spurred a new line of research that aims at introducing additional constraints that reflect more accurately the reality of the decision making process. *Bandits with Knapsacks* (BwK), a model formulated in its most general form in Badanidiyuru et al. (2013), fits into this framework and is characterized by the consumption of a limited supply of resources (e.g. time, money, and natural resources) that comes with every decision. This extension is motivated by a number of applications in electronic markets such as dynamic pricing with limited supply, see Besbes and Zeevi (2012), Besbes and Zeevi (2009), and Babaioff et al. (2012), online advertising Slivkins (2013), online bid optimization for sponsored search auctions Tran-Thanh et al. (2014a), and crowdsourcing Tran-Thanh et al. (2014b). A unifying paradigm of online learning is to evaluate algorithms based on their regret performance. In the BwK theory, this performance criterion is expressed as the gap between the total payoff of an optimal oracle algorithm aware of how the rewards and the amounts of resource consumption are generated and the total payoff of

the algorithm. Many approaches have been proposed to tackle the original MAB problem, where time is the only limited resource with a prescribed time horizon T , and the optimal regret bounds are now well documented. They can be classified into two categories with qualitatively different asymptotic growth rates. Many algorithms, such as UCB1 [Auer et al. \(2002a\)](#), Thomsom sampling [Agrawal and Goyal \(2012\)](#), and ϵ -greedy [Auer et al. \(2002a\)](#), achieve distribution-dependent, i.e. with constant factors that depend on the underlying unobserved distributions, asymptotic bounds on regret of order $\Theta(\ln(T))$, which is shown to be optimal in [Lai and Robbins \(1985\)](#). While these results prove very satisfying in many settings, the downside is that the bounds can get arbitrarily large if a malicious opponent was to select the underlying distributions in an adversarial fashion. In contrast, algorithms such as Exp3, designed in [Auer et al. \(2002b\)](#), achieve distribution-free bounds that can be computed in an online fashion, at the price of a less attractive growth rate $\Theta(\sqrt{T})$. The BwK theory lacks this clear-cut distinction. While provably optimal distribution-free bounds have recently been established, see [Agrawal and Devanur \(2014\)](#) and [Badanidiyuru et al. \(2013\)](#), there has been little progress toward the development of asymptotically optimal distribution-dependent regret bounds. To bridge the gap, in this paper we introduce algorithms with proven regret bounds which are asymptotically logarithmic in the initial supply of each resource, in three important cases that cover a wide range of applications:

- Case 1, where there is a single limited resource other than time, which is not limited, and the amount of resource consumed as a result of making a decision is stochastic. Applications in online advertising [Tran-Thanh et al. \(2010\)](#) and wireless sensor networks [Tran-Thanh et al. \(2012\)](#) fit in this framework;
- Case 2, where there are two limited resources, one of which is assumed to be time while the consumption of the other is stochastic, under a nondegenerate condition. Typical applications include online bid optimization for sponsored search auctions [Tran-Thanh et al. \(2014a\)](#), dynamic pricing with limited supply [Babaioff et al. \(2012\)](#), and dynamic procurement [Badanidiyuru et al. \(2013\)](#);
- Case 3, where there are arbitrarily many resources and the amounts of resources consumed as a result of making a decision are deterministic. Some applications to crowdsourcing markets [Tran-Thanh et al. \(2014b\)](#), ad allocation problems for a cost-per-impression pricing model [Slivkins \(2013\)](#), prediction with expert advice problems when querying an expert incurs some known cost [Amin et al. \(2015\)](#), and applications to the design of medical trials where the cost of a treatment is known ahead time fit in this framework.

1.2. Problem statement

At each time period $t \in \mathbb{N}$, a decision needs to be made among a predefined finite set of actions, represented by arms and labeled $k = 1, \dots, K$. We denote by a_t the arm pulled at time t . Pulling arm k at time t yields a random reward $r_{k,t} \in [0, 1]$ and incurs the consumption of $C \in \mathbb{N}$ different resource types by random amounts $c_{k,t}(1), \dots, c_{k,t}(C) \in [0, 1]^C$. Note that time itself may or may not be a limited resource. At any time t and for any arm k , the vector $(r_{k,t}, c_{k,t}(1), \dots, c_{k,t}(C))$ is jointly drawn from a fixed probability distribution ν_k independently from the past. The rewards and the amounts of resource consumption can be arbitrarily correlated across arms. We denote by $(\mathcal{F}_t)_{t \in \mathbb{N}}$ the natural filtration generated by the rewards and the amounts of resource consumption revealed to the decision maker, i.e. $((r_{a_t,t}, c_{a_t,t}(1), \dots, c_{a_t,t}(C)))_{t \in \mathbb{N}}$. The consumption of any

resource $i \in \{1, \dots, C\}$ is constrained by an initial budget $B(i) \in \mathbb{R}_+$. As a result, the decision maker can keep pulling arms only so long as he does not run out of any of the C resources and the game ends at time period τ^* , defined as:

$$\tau^* = \min\{t \in \mathbb{N} \mid \exists i \in \{1, \dots, C\}, \sum_{\tau=1}^t c_{a_\tau, \tau}(i) > B(i)\}. \quad (1)$$

Note that τ^* is a stopping time with respect to $(\mathcal{F}_t)_{t \geq 1}$. When it comes to choosing which arm to pull next, the difficulty for the decision maker lies in the fact that none of the underlying distributions, i.e. $(\nu_k)_{k=1, \dots, K}$, are initially known. Furthermore, the only feedback provided to the decision maker upon pulling arm a_t (but prior to selecting a_{t+1}) is $(r_{a_t, t}, c_{a_t, t}(1), \dots, c_{a_t, t}(C))$, i.e. the decision maker does not observe the rewards that would have been obtained and the amounts of resources that would have been consumed as a result of pulling a different arm. The goal is to design a non-anticipating algorithm that, at any time t , selects a_t based on the information acquired in the past so as to keep the pseudo regret defined as:

$$R_{B(1), \dots, B(C)} = \text{ER}_{\text{OPT}}(B(1), \dots, B(C)) - \mathbb{E}\left[\sum_{t=1}^{\tau^*-1} r_{a_t, t}\right], \quad (2)$$

as small as possible, where $\text{ER}_{\text{OPT}}(B(1), \dots, B(C))$ is the maximum expected sum of rewards that can be obtained by a non-anticipating oracle algorithm that has knowledge of the underlying distributions. Here, an algorithm is said to be non-anticipating if the decision to pull a given arm does not depend on the future observations. We develop algorithms and establish distribution-dependent regret bounds, that hold for any choice of the unobserved underlying distributions $(\nu_k)_{k=1, \dots, K}$, as well as distribution-independent regret bounds. This entails studying the asymptotic behavior of $R_{B(1), \dots, B(C)}$ when all the budgets $(B(i))_{i=1, \dots, C}$ go to infinity. In order to simplify the analysis, it is convenient to assume that the ratios $(B(i)/B(C))_{i=1, \dots, C}$ are constants independent of any other relevant quantities and to denote $B(C)$ by B .

Assumption 1 *For any resource $i \in \{1, \dots, C\}$, we have $B(i) = b(i) \cdot B$ for some fixed constant $b(i) \in (0, 1]$. Hence $b = \min_{i=1, \dots, C} b(i)$ is a positive quantity.*

When time is a limited resource, we use the notation T in place of B . Assumption 1 is only necessary when deriving distribution-dependent regret bounds. This assumption is widely used in the dynamic pricing literature where the inventory scales linearly with the time horizon, see [Besbes and Zeevi \(2012\)](#) and [Johnson et al. \(2015\)](#).

As the mean turns out to be an important statistics, we denote the mean reward and amounts of resource consumption by $\mu_k^r, \mu_k^c(1), \dots, \mu_k^c(C)$ and their respective empirical estimates by $\bar{r}_{k, t}, \bar{c}_{k, t}(1), \dots, \bar{c}_{k, t}(C)$. These estimates depend on the number of times each arm has been pulled by the decision maker up to, but not including, time t , which we write $n_{k, t}$. We end with a general assumption that we use throughout the paper.

Assumption 2 *There exists $\epsilon > 0$ such that, for any resource $i \in \{1, \dots, C\}$ and for any arm $k \in \{1, \dots, K\}$, $\mu_k^c(i) \geq \epsilon$. Moreover ϵ is assumed to be known to the decision maker prior to starting the game.*

The first part of Assumption 2 is meant to have the game end in finite time almost surely. Strictly speaking, we only need the weaker assumption that pulling any arm incurs the consumption of at least one resource. We use the stronger Assumption 2 to simplify the presentation but all the proofs can be easily adapted to accommodate the weaker version. The second part of Assumption 2 serves two purposes. First, the analysis conducted under this additional assumption can be extended to the general setting at the price of more technicalities and the loss of finite-time regret bounds, as detailed in Section 6. Second, this leads to better bounds on regret if the decision maker happens to have access to such information.

1.3. Contributions

We design a general-purpose algorithm running in time polynomial in K for which we establish:

- (a) a $O(K \cdot \ln(B))$ (resp. $O(\sqrt{K \cdot B \cdot \ln(B)})$) distribution-dependent (resp. distribution-free) regret bound when $C = 1$;
- (b) a $O(K^2 \cdot \ln(T))$ (resp. $O(K \cdot \sqrt{T \cdot \ln(T)})$) distribution-dependent (resp. distribution-free) regret bound when: (i) $C = 2$, (ii) resource $i = 2$ is time, (iii) the following nondegenerate condition holds:

$$|\mu_k^c(1) - b| \geq \epsilon, \quad k = 1, \dots, K;$$

- (c) a $O(K^C \cdot \ln(B))$ (resp. $O(\sqrt{K^C \cdot B \cdot \ln(B)})$) distribution-dependent (resp. distribution-free) regret bound when $c_{k,t}(1), \dots, c_{k,t}(C)$ are deterministic for any time t and any arm k . By tweaking the algorithm, we can also get a $O(K \cdot \ln(B))$ distribution-dependent regret bound but no non-trivial distribution-free regret bound, and the algorithm is also time polynomial in C in this case.

All of the aforementioned regret bounds also hold up to logarithmic factors when ϵ is not known to the decision maker prior to starting the game.

1.4. Literature Review

The Bandits with Knapsacks framework was first introduced in its full generality in [Badanidiyuru et al. \(2013\)](#), but special cases had been studied before, see for example [Tran-Thanh et al. \(2010\)](#), [Tran-Thanh et al. \(2012\)](#), [Ding et al. \(2013\)](#), and [Babaioff et al. \(2012\)](#). Since the standard MAB problem fits in the BwK framework, with time being the only scarce resource, the results listed in the introduction tend to suggest that regret bounds with logarithmic growth with respect to the budgets may be possible for BwK problems but very few such results are documented. When there are arbitrarily many resources and a time horizon, [Badanidiyuru et al. \(2013\)](#) and [Agrawal and Devanur \(2014\)](#) obtain $\tilde{O}(\sqrt{K \cdot T})$ distribution-free bounds on regret that hold on average as well as with high probability, where the \tilde{O} notation hides logarithmic factors. These results were later extended to the contextual version of the problem in [Badanidiyuru et al. \(2014\)](#) and [Agrawal et al. \(2016\)](#). [Johnson et al. \(2015\)](#) extend Thompson sampling to tackle the general BwK problem and obtain distribution-dependent bounds on regret of order $\tilde{O}(\sqrt{T})$, with an unspecified dependence on K , under a nondegenerate condition when one of the limited resources is time. [Combes et al. \(2015\)](#) consider a closely related framework that allows to model any history-dependent constraint on the number of times any arm can be pulled along with a time horizon T and obtain $O(K \cdot \ln(T))$ regret bounds. However, the benchmark oracle algorithm they use to define regret is much weaker than the one considered here as

it only has knowledge of the distributions of the rewards, as opposed to the joint distributions of the rewards and the amounts of resource consumption. [Babaioff et al. \(2012\)](#) establish a $\Omega(\sqrt{T})$ distribution-dependent lower bound on regret for a dynamic pricing problem which can be cast as a BWK problem with a time horizon, a stochastic resource, and a continuum of arms. This lower bound does not apply here as we are considering finitely many arms and it is well known that there is an exponential separation between the best possible expected regret when we move from finitely many arms to uncountably many arms for the standard MAB problem, see [Kleinberg and Leighton \(2003\)](#). [Tran-Thanh et al. \(2012\)](#) tackles BWK problems with a single limited resource whose consumption is deterministic and constrained by a global budget B and obtain $O(K \cdot \ln(B))$ regret bounds. This result was later extended to the case of a stochastic resource in [Xia et al. \(2015\)](#). [Wu et al. \(2015\)](#) study a contextual version of the BWK problem when there are two limited resources, one of which is assumed to be time while the consumption of the other is deterministic, and obtain $O(K \cdot \ln(T))$ regret bounds under a nondegenerate condition.

Organization. The remainder of the paper is organized as follows. We present the algorithmic ideas underlying our approach in Section 2 and apply these ideas to Cases (a), (b), and (c) in Sections 3, 5, and 4 respectively. We relax some of the assumptions made in the course of proving the regret bounds and discuss extensions in Section 6. All the proofs are deferred to the Appendix.

2. Algorithmic ideas

2.1. Preliminaries

To handle the exploration-exploitation trade-off, an approach that has proved to be particularly successful hinges on the *optimism in the face of uncertainty* paradigm. The idea is to consider all plausible scenarios consistent with the information collected so far and to select the decision that yields the most revenue among all the scenarios identified. Concentration inequalities are intrinsic to the paradigm as they enable the development of systematic closed form confidence intervals on the quantities of interest, which together define a set of plausible scenarios. We make repeated use of the following result.

Lemma 1 *Hoeffding's inequality, [Hoeffding \(1963\)](#)*

Consider X_1, \dots, X_n n random variables with support in $[0, 1]$.

If $\forall t \leq n \mathbb{E}[X_t \mid X_1, \dots, X_{t-1}] \leq \mu$, then $\mathbb{P}[X_1 + \dots + X_n \geq n\mu + a] \leq \exp(-\frac{2a^2}{n}) \quad \forall a \geq 0$.
If $\forall t \leq n \mathbb{E}[X_t \mid X_1, \dots, X_{t-1}] \geq \mu$, then $\mathbb{P}[X_1 + \dots + X_n \leq n\mu - a] \leq \exp(-\frac{2a^2}{n}) \quad \forall a \geq 0$.

[Auer et al. \(2002a\)](#) follow the *optimism in the face of uncertainty* paradigm to develop the Upper Confidence Bound algorithm (UCB1). UCB1 is based on the following observations: (i) the optimal strategy always consists in pulling the arm with the highest mean reward when time is the only limited resource, (ii) informally, Lemma 1 shows that $\mu_k^r \in [\bar{r}_{k,t} - \epsilon_{k,t}, \bar{r}_{k,t} + \epsilon_{k,t}]$ at time t with probability at least $1 - \frac{2}{t^3}$ for $\epsilon_{k,t} = \sqrt{\frac{2 \cdot \ln(t)}{n_{k,t}}}$, irrespective of the number of times arm k has been pulled. Based on these observations, UCB1 always selects the arm with highest UCB index, i.e. $a_t \in \operatorname{argmax}_{k=1, \dots, K} I_{k,t}$, where the UCB index of arm k at time t is defined as $I_{k,t} = \bar{r}_{k,t} + \epsilon_{k,t}$. The first term can be interpreted as an exploitation term, the ultimate goal being to maximize revenue, while the second term is an exploration term, the smaller $n_{k,t}$, the bigger it is. This fruitful paradigm goes well beyond this special case and many extensions of

UCB1 have been designed to tackle variants of the MAB problem, see for example [Slivkins \(2013\)](#). [Agrawal and Devanur \(2014\)](#) embrace the same ideas to tackle BwK problems. The situation is more complex in this all-encompassing framework as the optimal oracle algorithm involves pulling several arms. In fact, finding the optimal pulling strategy given the knowledge of the underlying distributions is already a challenge in its known, see [Papadimitriou and Tsitsiklis \(1999\)](#) for a study of the computational complexity of similar problems. This raises the question of how to evaluate $\text{ER}_{\text{OPT}}(B(1), \dots, B(C))$ in (2). To overcome this issue, [Badanidiyuru et al. \(2013\)](#) upper bound the total expected payoff of any non-anticipating algorithm by the optimal value of a linear program, which is easier to compute.

Lemma 2 *Adapted from [Badanidiyuru et al. \(2013\)](#)*

The total expected payoff of any non-anticipating algorithm is no greater than B times the optimal value of the linear program:

$$\begin{aligned} \sup_{(\xi_k)_{k=1, \dots, K}} \quad & \sum_{k=1}^K \mu_k^r \cdot \xi_k \\ \text{subject to} \quad & \sum_{k=1}^K \mu_k^c(i) \cdot \xi_k \leq b(i), \quad i = 1, \dots, C \\ & \xi_k \geq 0, \quad k = 1, \dots, K \end{aligned} \tag{3}$$

plus the constant term $\frac{1}{\epsilon}$.

The optimization problem (3) can be interpreted as follows. For any arm k , $B \cdot \xi_k$ corresponds to the expected number of times arm k is pulled by the optimal algorithm. Hence, assuming we introduce a dummy arm 0 which is equivalent to skipping the current round, ξ_k corresponds to the probability of pulling arm k at any round when there is a time horizon T . Observe that the constraints restrict the feasible set of expected number of pulls by imposing that the amounts of resources consumed be no greater than their respective budgets in expectations, as opposed to almost surely which would be a more stringent constraint. This explains why the optimal value of (3) is larger than the maximum achievable payoff. In this paper, we use standard linear programming notions such as the concept of a basis or of a basic feasible solution. We refer to [Bertsimas and Tsitsiklis \(1997\)](#) for an introduction to linear programming. For x a feasible basis for (3), we denote by $(\xi_k^x)_{k=1, \dots, K}$ the corresponding basic feasible solution and by $\text{obj}_x = \sum_{k=1}^K \xi_k^x \cdot \mu_k^r$ its objective function. From Lemma 2, we derive:

$$R_{B(1), \dots, B(C)} \leq B \cdot \text{obj}_{x^*} - \mathbb{E} \left[\sum_{t=1}^{\tau^*} r_{a_t, t} \right] + O(1), \tag{4}$$

where x^* is an optimal basis for (3). For mathematical convenience, we consider that the game carries on even if one of the resources is already exhausted so that a_t is well-defined for any $t \in \mathbb{N}$. Of course, the rewards obtained for $t \geq \tau^*$ are not taken account of in the decision maker's payoff when establishing regret bounds.

2.2. Solution methodology

Lemma 2 also provides insight into designing algorithms. The idea is to incorporate confidence intervals on the mean rewards and the mean amounts of resource consumption into the offline optimization problem (3) and to base the decision upon the resulting optimal solution. There are several

ways to carry out this task, each leading to a different algorithm. When there is a time horizon T , [Agrawal and Devanur \(2014\)](#) use high-probability lower (resp. upper) bounds on the mean amounts of resource consumption (resp. rewards) in place of the unknown mean values in (3) and pull an arm at random according to the resulting optimal distribution. Specifically, at any round t , they compute an optimal solution $(\xi_{k,t}^*)_{k=1,\dots,K}$ to the linear program:

$$\begin{aligned} & \sup_{(\xi_k)_{k=1,\dots,K}} \sum_{k=1}^K (\bar{r}_{k,t} + \epsilon_{k,t}) \cdot \xi_k \\ & \text{subject to} \quad \sum_{k=1}^K (\bar{c}_{k,t}(i) - \epsilon_{k,t}) \cdot \xi_k \leq (1 - \gamma) \cdot b(i), \quad i = 1, \dots, C \\ & \quad \sum_{k=1}^K \xi_k \leq 1 \\ & \quad \xi_k \geq 0, \quad k = 1, \dots, K, \end{aligned} \tag{5}$$

for a well-chosen $\gamma \in (0, 1)$, and pull arm k with probability $\xi_{k,t}^*$ or skip the round with probability $1 - \sum_{k=1}^K \xi_{k,t}^*$. If we relate this approach to UCB1, the intuition is clear: the idea is to be optimistic about both the rewards and the amounts of resource consumption. We argue that this approach cannot yield logarithmic regret bounds. First, because γ has to be of order $1/\sqrt{T}$. Secondly, because, even if we were given an optimal distribution $(\xi_k^*)_{k=1,\dots,K}$ solution to (3) prior to starting the game, consistently choosing which arm to pull at random according to this distribution at every round would incur regret $\Omega(\sqrt{T})$, as we next show.

Lemma 3 *Pulling arm k with probability ξ_k^* at any round t yields a regret of order $\Omega(\sqrt{T})$ unless pulling any arm in $\text{supp}(x^*)$ incurs the deterministic consumption of the same amounts of resources.*

The fundamental shortcoming of this approach is that it systematically leads us to plan to consume the same average amount of resources per round $b(i)$, for each resource $i = 1, \dots, C$, irrespective of whether we have significantly over- or under-consumed in the past. To address this issue, we propose the following template algorithm also based on the linear relaxation (3).

UCB-Simplex *Take $\lambda \geq 1$ (λ will need to be carefully chosen). The algorithm is preceded by an initialization phase which consists in pulling each arm a given number of times, to be specified. At each subsequent time period t , proceed as follows.*

Step 1: *Find an optimal basis x_t to the linear program:*

$$\begin{aligned} & \sup_{(\xi_k)_{k=1,\dots,K}} \sum_{k=1}^K (\bar{r}_{k,t} + \lambda \cdot \epsilon_{k,t}) \cdot \xi_k \\ & \text{subject to} \quad \sum_{k=1}^K \bar{c}_{k,t}(i) \cdot \xi_k \leq b(i), \quad i = 1, \dots, C \\ & \quad \xi_k \geq 0, \quad k = 1, \dots, K \end{aligned} \tag{6}$$

We denote the corresponding basic feasible solution by $(\xi_{k,t}^{x_t})_{k=1,\dots,K}$.

Step 2: *Identify the arms involved in the optimal basis, i.e. $\text{supp}(x_t) = \{k \in \{1, \dots, K\} \mid \xi_{k,t}^{x_t} >$*

0}. There are at most $\min(K, C)$ such arms. Use a load balancing algorithm \mathcal{A}_{x_t} , to be specified, to determine which of these arms to pull.

The Simplex algorithm is an obvious choice to carry out Step 1, especially because we only have to update one column of the constraint matrix per round for (6), as opposed to a $K \times C$ submatrix for (5), which makes its warm-starting properties attractive. However, note that this can also be done in time polynomial in K and C , see Grötschel et al. (2012). If we compare (6) with (5), the idea remains to be overly optimistic, but only about the rewards, thus transferring the burden of exploration from the constraints to the objective function through the exploration factor λ . The details of Step 2 are purposefully left out and will be specified for each of the cases treated in this paper. When there is a time horizon T , the general idea is to determine, at any time period t and for each resource $i = 1, \dots, C$, whether we have over- or under-consumed in the past and to perturb the probability distribution $(\xi_{k,t}^{x_t})_{k=1,\dots,K}$ accordingly to get back on track.

The algorithm we propose is intrinsically tied to the existence of a basic feasible optimal solution to (3) and (6). We denote by \mathcal{B} (resp. \mathcal{B}_t) the set of feasible bases for (3) (resp. (6)). Step 1 of UCB-Simplex can be interpreted as an extension of the index-based decision rule of UCB1. Indeed, Step 1 consists in assigning an index $I_{x,t}$ to each basis $x \in \mathcal{B}_t$ and to select $x_t \in \operatorname{argmax}_{x \in \mathcal{B}_t} I_{x,t}$, where $I_{x,t} = \operatorname{obj}_{x,t} + E_{x,t}$ with a clear separation between the exploitation term, $\operatorname{obj}_{x,t} = \sum_{k=1}^K \xi_{k,t}^x \cdot \bar{r}_{k,t}$, and the exploration term, $E_{x,t} = \lambda \cdot \sum_{k=1}^K \xi_{k,t}^x \cdot \epsilon_{k,t}$. Observe that for $x \in \mathcal{B}_t$ that is also feasible for (3), $(\xi_{k,t}^x)_{k=1,\dots,K}$ and $\operatorname{obj}_{x,t}$ are plug-in estimates of $(\xi_k^x)_{k=1,\dots,K}$ and obj_x . Also note that when $\lambda = 1$ and when time is the only limited resource, UCB-Simplex is identical to UCB1 as Step 2 is unambiguous in this special case, each basis involving a single arm. For any $x \in \mathcal{B}$, we define $\Delta_x = \operatorname{obj}_{x^*} - \operatorname{obj}_x \geq 0$ as the optimality gap. A feasible basis x is said to be suboptimal if $\Delta_x > 0$. At any time t , $n_{x,t}$ denotes the number of times basis x has been selected at Step 1 up to time t while $n_{k,t}^x$ denotes the number of times arm k has been pulled up to time t when selecting x at Step 1. For all the cases treated in this paper, we will show that, under a nondegeneracy assumption, Step 1 of UCB-Simplex guarantees that a suboptimal basis cannot be selected more than $O(\ln(B))$ times on average, a result reminiscent of the regret analysis of UCB1 carried out in Auer et al. (2002a). However, in stark contrast with the situation of a single limited resource, this is merely a prerequisite to establish a $O(\ln(B))$ bound on regret. Indeed, a low regret algorithm must also balance the load between the arms as closely as possible to optimality. Hence, the choice of the load balancing algorithms \mathcal{A}_x is crucial to obtain $O(\ln(B))$ regret bounds.

3. A single limited resource

In this section, we tackle the case of a single resource whose consumption is limited by a global budget B , i.e. $C = 1$ and $b(1) = 1$. To simplify the notations, we omit the indices identifying the resources as there is only one, i.e. we write μ_k^c , $c_{k,t}$, and $\bar{c}_{k,t}$ as opposed to $\mu_k^c(1)$, $c_{k,t}(1)$, and $\bar{c}_{k,t}(1)$.

Specification of the algorithm. We implement UCB-Simplex with $\lambda = 1 + \frac{1}{\epsilon}$. The initialization step consists in pulling each arm until the amount of resource consumed as a result of pulling that arm is non-zero. The purpose of this step is to have $\bar{c}_{k,t} > 0$ for all periods to come and for all arms. Step 2 of UCB-Simplex is unambiguous here as basic feasible solutions involve a single arm. Hence, we identify a basis $x = \{k\}$ with the corresponding arm and write $x = k$ to simplify the notations. In particular, $k^* \in \{1, \dots, K\}$ identifies an optimal arm in the sense defined in Section

2. For any arm k , the exploration and exploitation terms defined in Section 2 specialize to:

$$\text{obj}_{k,t} = \frac{\bar{r}_{k,t}}{\bar{c}_{k,t}} \text{ and } E_{k,t} = (1 + \frac{1}{\epsilon}) \cdot \frac{\epsilon_{k,t}}{\bar{c}_{k,t}},$$

while $\text{obj}_k = \frac{\mu_k^r}{\mu_k^c}$, so that:

$$k^* \in \operatorname{argmax}_{k=1,\dots,K} \frac{\mu_k^r}{\mu_k^c}, \quad a_t \in \operatorname{argmax}_{k=1,\dots,K} \frac{\bar{r}_{k,t} + (1 + \frac{1}{\epsilon}) \cdot \epsilon_{k,t}}{\bar{c}_{k,t}}, \text{ and } \Delta_k = \frac{\mu_{k^*}^r}{\mu_{k^*}^c} - \frac{\mu_k^r}{\mu_k^c}.$$

We point out that, for the particular setting considered in this section, UCB-Simplex is almost identical to the fractional KUBE algorithm proposed in [Tran-Thanh et al. \(2012\)](#) to tackle the case of a single resource whose consumption is deterministic. It only differs by the presence of the scaling factor $1 + \frac{1}{\epsilon}$ to favor exploration over exploitation, which becomes unnecessary when the amounts of resource consumed are deterministic, see Section 6.

Regret analysis. We omit the initialization step in the theoretical analysis because the amount of resource consumed is $O(1)$ and the reward obtained is non-negative and not taken account of in the decision maker's total payoff. Moreover, the initialization step ends in finite time almost surely as a result of Assumption 2. First observe that (4) specializes to:

$$R_B \leq B \cdot \frac{\mu_{k^*}^r}{\mu_{k^*}^c} - \mathbb{E} \left[\sum_{t=1}^{\tau^*} r_{a_t,t} \right] + O(1). \quad (7)$$

To bound the right-hand side, we start by estimating the expected time horizon.

Lemma 4 *For any non-anticipating algorithm, we have: $\mathbb{E}[\tau^*] \leq \frac{B+1}{\epsilon}$.*

The next result is crucial. Used in combination with Lemma 4, it shows that any suboptimal arm is pulled at most $O(\ln(B))$ times in expectations, a well-known result for UCB1, see [Auer et al. \(2002a\)](#). The proof is along the same lines as the proof for UCB1, namely we assume that arm k has already been pulled more than $\Theta(\ln(\tau^*)/(\Delta_k)^2)$ times and conclude that arm k cannot be pulled more than a few more times, with the additional difficulty of having to deal with the random stopping time and the fact that the amount of resource consumed at each step is stochastic.

Lemma 5 *For any suboptimal arm k , we have:*

$$\mathbb{E}[n_{k,\tau^*}] \leq \left(\frac{4}{\epsilon}\right)^4 \cdot \frac{\mathbb{E}[\ln(\tau^*)]}{(\Delta_k)^2} + \frac{4\pi^2}{3\epsilon^2}.$$

Building on the last two results, we recover the result of [Xia et al. \(2015\)](#) with a finite-time regret bound which generalizes the one obtained by [Auer et al. \(2002a\)](#) when time is the only scarce resource.

Proposition 6

$$R_B \leq \left(\frac{4}{\epsilon}\right)^4 \cdot \left(\sum_{k \mid \Delta_k > 0} \frac{1}{\Delta_k} \right) \cdot \ln\left(\frac{B+1}{\epsilon}\right) + O(1).$$

Observe that the set of optimal arms, namely $\operatorname{argmax}_k \mu_k^r / \mu_k^c$, does not depend on B and that $\Delta_k = \mu_{k^*}^r / \mu_{k^*}^c - \mu_k^r / \mu_k^c$ is a constant independent of B for any suboptimal arm. We conclude that $R_B = O(\frac{K}{\Delta} \cdot \ln(B))$ with $\Delta = \min_k | \Delta_k > 0 | \Delta_k$. Interestingly, the algorithm we propose does not rely on B to achieve this regret bound, much like what happens for UCB1 with the time horizon, see [Auer et al. \(2002a\)](#). This result is optimal up to constant factors as the standard MAB problem is a special case of the framework considered in this section, see [Lai and Robbins \(1985\)](#) for a proof of a lower bound in this context. Building on Proposition 6, we can also derive a near-optimal distribution-free regret bound in the same fashion as for UCB1.

Proposition 7

$$R_B \leq \left(\frac{4}{\epsilon}\right)^2 \cdot \sqrt{K \cdot \frac{B+1}{\epsilon} \cdot \ln\left(\frac{B+1}{\epsilon}\right)} + O(1).$$

We conclude that $R_B = O(\sqrt{K \cdot B \cdot \ln(B)})$, where the hidden factors are independent of the underlying distributions $(\nu_k)_{k=1, \dots, K}$.

4. Arbitrarily many limited resources whose consumption are deterministic

In this section, we study the case of multiple limited resources when the amounts of resources consumed as a result of pulling an arm are deterministic and globally constrained by prescribed budgets $(B(i))_{i=1, \dots, C}$, where C is the number of resources. Because the amounts of resources consumed are deterministic, the exploration (resp. exploitation) terms defined in Section 2 specialize to $\operatorname{obj}_{x,t} = \sum_{k=1}^K \xi_k^x \cdot \bar{r}_{k,t}$ (resp. $E_{x,t} = \sum_{k=1}^K \xi_k^x \cdot \epsilon_{k,t}$) and we can substitute the notation $\mu_k^c(i)$ for $c_k(i)$. We point out that the stopping time need not be deterministic as the decision to select an arm is based on the past realizations of the rewards. We define $r_{1, \dots, C} \leq \min(C, K)$ as the rank of the matrix $(c_k(i))_{1 \leq k \leq K, 1 \leq i \leq C}$.

Specification of the algorithm. We implement UCB-Simplex with an initialization step which now consists in pulling each arm at least $r_{1, \dots, C}$ times. The motivation behind this step is mainly technical and is simply meant to have:

$$n_{k,t} \geq r_{1, \dots, C} + \sum_{x \in \mathcal{B} \mid k \in \operatorname{supp}(x)} n_{k,t}^x \quad \forall t, \forall k \in \{1, \dots, K\}. \quad (8)$$

Compared to Section 3, we choose to take $\lambda = 1$ and we are now required to specify the load balancing algorithms involved in Step 2 of UCB-Simplex as the feasible bases selected at Step 1 may involve several arms. Although Step 2 will also need to be specified in Section 5, designing good load balancing algorithms is arguably easier when the amounts of resources consumed as a result of pulling arms are deterministic because the optimal load balance is known for each basis from the start. Nonetheless, one challenge remains: we can never identify the (possibly many) optimal bases of (3) with absolute certainty. As a result, every basis selected at Step 1 should be treated as potentially optimal when balancing the load between the arms involved in this basis, but this inevitably causes some interference issues as an arm may be involved in several bases, and worst, possibly several optimal bases. Therefore, one point that will appear to be of particular importance in the analysis is the use of load balancing algorithms that are decoupled from one another, in the sense that they do not rely on what happened when selecting other arms. More specifically, we use the following class of load balancing algorithms.

Load balancing algorithm \mathcal{A}_x for a feasible basis $x \in \mathcal{B}$

If basis x is selected at time t , pull any arm $k \in \text{supp}(x)$ such that $n_{k,t}^x \leq n_{x,t} \cdot \frac{\xi_k^x}{\sum_{l=1}^K \xi_l^x}$.

The load balancing algorithms \mathcal{A}_x thus defined are decoupled because, for each basis, the number of times an arm has been pulled when selecting another basis is not taken account of. The following lemma shows that \mathcal{A}_x is always well-defined and guarantees that the ratios $(n_{k,t}^x/n_{l,t}^x)_{k,l \in \text{supp}(x)}$ remain close to the optimal ones $(\xi_k^x/\xi_l^x)_{k,l \in \text{supp}(x)}$ at all times.

Lemma 8 \mathcal{A}_x is always well-defined and moreover, at any time t , for any basis $x \in \mathcal{B}$, and for any arm $k \in \text{supp}(x)$:

$$n_{x,t} \cdot \frac{\xi_k^x}{\sum_{l=1}^K \xi_l^x} - r_{1,\dots,C} \leq n_{k,t}^x \leq n_{x,t} \cdot \frac{\xi_k^x}{\sum_{l=1}^K \xi_l^x} + 1 \text{ almost surely,}$$

while $n_{k,t}^x = 0$ for any arm $k \notin \text{supp}(x)$.

Observe that implementing the load balancing algorithms \mathcal{A}_x may require a memory storage capacity exponential in C and polynomial in the number of arms, although always bounded by $O(B)$ (because we do not need to keep track of $n_{k,t}^x$ if x has never been selected). In practice, only a few bases will be selected at Step 1, so that a hash table is an appropriate data structure to store the sequences $(n_{k,t}^x)_{k \in \text{supp}(x)}$. In Section 6, we introduce another class of load balancing algorithms that is both time and memory efficient while still guaranteeing $O(\ln(B))$ regret bounds (under an additional assumption) but no distribution-free regret bounds.

Regret Analysis. We discard the initialization step in the theoretical study because the amounts of resources consumed are bounded by a constant and the total reward obtained is non-negative and not taken account of in the decision maker's total payoff. We start by estimating the expected time horizon.

Lemma 9 For any non-anticipating algorithm, we have: $\mathbb{E}[\tau^*] \leq \frac{B+1}{\epsilon}$.

We follow by bounding the number of times any suboptimal basis can be selected at Step 1, in the same spirit as in Section 3.

Lemma 10 For any suboptimal basis $x \in \mathcal{B}$, we have:

$$\mathbb{E}[n_{x,\tau^*}] \leq \frac{16 \cdot r_{1,\dots,C}}{\epsilon^2} \cdot \frac{\mathbb{E}[\ln(\tau^*)]}{(\Delta_x)^2} + r_{1,\dots,C} \cdot \frac{\pi^2}{3}.$$

Lemma 10 used in combination with Lemma 9 shows that a suboptimal basis is selected at most $O(\ln(B))$ times. To establish the regret bound, it remains to lower bound the expected total payoff derived when selecting any of the optimal bases. This is more involved than in the case of a single limited resource because the load balancing step comes into play at this stage.

Proposition 11

$$R_{B(1),\dots,B(C)} \leq \frac{16 \cdot r_{1,\dots,C}}{b \cdot \epsilon^2} \cdot \left(\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{1}{\Delta_x} \right) \cdot \ln\left(\frac{B+1}{\epsilon}\right) + O(1).$$

Since the number of feasible bases for the linear program (3) is at most $\binom{K+r_1, \dots, C}{K}$, we get the distribution-dependent regret bound $O\left(\binom{K+r_1, \dots, C}{K} \cdot \frac{\ln(B)}{\Delta}\right)$ where $\Delta = \min_{x \in \mathcal{B} \mid \Delta_x > 0} \Delta_x$. We point out that, under the assumption that there is a unique optimal basis to (3), the alternative class of load balancing algorithms introduced in Section 6 yields a better dependence on K and C with a regret bound of $O\left(K \cdot \frac{\ln(B)}{\Delta^2}\right)$. Along the same lines as in Section 3, the distribution-dependent bound of Proposition 11 almost immediately implies a distribution-free one.

Proposition 12

$$R_{B(1), \dots, B(C)} \leq \frac{4}{b \cdot \epsilon} \cdot \sqrt{r_{1, \dots, C} \cdot |\mathcal{B}| \cdot \frac{B+1}{\epsilon} \cdot \ln\left(\frac{B+1}{\epsilon}\right)} + O(1).$$

We conclude that $R_{B(1), \dots, B(C)} = O\left(\sqrt{r_{1, \dots, C} \cdot \binom{K+r_1, \dots, C}{K} \cdot B \cdot \ln(B)}\right)$, where the hidden factors are independent of the underlying distributions $(\nu_k)_{k=1, \dots, K}$. We stress that the dependence on K and C is not optimal since Agrawal and Devanur (2014) obtain a $\tilde{O}(\sqrt{K \cdot B})$ bound on regret, where the \tilde{O} notation hides factors logarithmic in B .

5. A time horizon and another limited resource

In this section, we investigate the case of two limited resources, one of which is assumed to be time, with a time horizon T , while the consumption of the other is stochastic and constrained by a global budget B . To simplify the notations, we omit the indices identifying the resources since the second limited resource is time and we write μ_k^c , $c_{k,t}$, $\bar{c}_{k,t}$, B , and T as opposed to $\mu_k^c(1)$, $c_{k,t}(1)$, $\bar{c}_{k,t}(1)$, $B(1)$, and $B(2)$. Moreover, we refer to resource $i = 1$ as “the” resource. Observe that, in the particular setting considered in this section, $\tau^* = \min(\tau(B), T + 1)$ with $\tau(B) = \min\{t \in \mathbb{N} \mid \sum_{\tau=1}^t c_{a_\tau, \tau} > B\}$. Note that the budget constraint is not limiting if $B \geq T$, in which case the problem reduces to the standard MAB problem. Hence, without loss of generality under Assumption 1, we assume that the budget scales linearly with time, i.e. $B = b \cdot T$ for a fixed constant $b \in (0, 1)$, and we study the asymptotic regime $T \rightarrow \infty$.

Nondegeneracy assumption. We point out that, in degenerate scenarios, using the linear relaxation (3) as an upper bound on the the maximum achievable payoff $\text{ER}_{\text{OPT}}(B, T)$ already dooms us to $\Omega(\sqrt{T})$ regret bounds. Precisely, if there exists an arm k^* such that $\{k^*\}$ is the unique optimal basis for (3) and $\mu_{k^*}^c = b$, then $T \cdot \text{obj}_{\{k^*\}} = T \cdot \mu_{k^*}^r \geq \text{ER}_{\text{OPT}}(B, T) + \Omega(\sqrt{T})$. This is because consistently pulling arm k^* is nothing but an i.i.d. strategy which implies, along the same lines as in Lemma 3, that $\mathbb{E}[\tau^*] = T - \Omega(\sqrt{T})$. Dealing with these degenerate scenarios thus calls for a completely approach than the one taken on in the BwK literature and we choose instead to rule them out by strengthening Assumption 2 in such a way that there can be no degenerate optimal basis for (3).

Assumption 3 *Assumption 2 holds and we also have $|\mu_k^c - b| \geq \epsilon$ for all arms $k \in \{1, \dots, K\}$.*

This assumption can be relaxed to some extent at the price of more technicalities and the loss of finite-time regret bounds, which turn into asymptotic ones. For example if Assumption 3 holds but ϵ is not known prior to starting the game, ϵ should be taken as a vanishing function of T along the same lines as what is done in Section 6 for the case of a single resource. However, the minimal

assumption is that there is no degenerate optimal basis for (3). As a final remark, we stress that Assumption 3 is only necessary to carry out the analysis but the algorithm can be implemented in any case.

Specification of the algorithm. We implement UCB-Simplex with $\lambda = 1 + \frac{1}{\epsilon}$. Because the amount of resource consumed at each time step is a random variable, a feasible basis for (6) may not be feasible for (3) and conversely. This is in contrast to the situation studied in Section 4. As a consequence, x^* may not be feasible for (6), thus effectively preventing it from being selected at Step 1 of UCB-Simplex, and a basis infeasible for (3) may be selected instead. To guarantee that these events, however still possible, occur with low probability, the initialization phase consists in pulling each arm $\lceil \frac{1}{\epsilon^2} \ln(T) \rceil$ times. Hence, we start implementing Steps 1 and 2 at round $t_i = K \cdot \lceil \frac{1}{\epsilon^2} \ln(T) \rceil + 1$ and, at any time $t \geq t_i$, any arm has been pulled at least $\frac{1}{\epsilon^2} \cdot \ln(t)$ times, i.e:

$$n_{k,t} \geq \frac{1}{\epsilon^2} \cdot \ln(t), \quad t = t_i, \dots, T, k = 1, \dots, K. \quad (9)$$

It turns out that this initialization step is not necessary to get logarithmic regret bounds because, at any time t , if an arm k has not been pulled more than $\ln(t)/\epsilon^2$ times, then basis $\{k\}$ will be optimal for (6) with high probability. While this argument can be made to work in a mathematically precise way, it significantly complicates the analysis and it does not improve the regret bounds. For these reasons, we choose to implement an initialization phase instead. We also need to specify Step 2 of UCB-Simplex because the basis selected at Step 1 may involve up to two arms. To simplify the notations, we introduce two dummy arms $k = 0$ and $k = K + 1$ with reward 0 and resource consumption 0 and 1 respectively so that every basis involves two arms. Specifically, for any arm such that $\mu_k^c < b$, the basis $\{k\}$ is mapped to the basis $\{k, K + 1\}$ and otherwise, if $\mu_k^c > b$, the basis $\{k\}$ is mapped to the basis $\{0, k\}$. As a result of this mapping, the inequality $\sum_k \mu_k^c \cdot \xi_k \leq b$ is binding for all bases that are feasible for (3). When we select basis $x_t = \{k, l\}$, we use a load balancing algorithm specific to this basis, which we recall is denoted by $\mathcal{A}_{\{k,l\}}$, to determine which of these two arms to pull. Similarly as in Section 4, using load balancing algorithms that are decoupled from one another is crucial because the decision maker can never identify the optimal bases with absolute certainty, which implies that each basis should be treated as potentially optimal when balancing the load between the arms, but this inevitably causes interference issues as an arm may be involved in several bases. Compared to Section 4, we face an additional challenge when designing the load balancing algorithms: the optimal load balances are initially unknown to the decision maker. It turns out that we can still approximately achieve the unknown optimal load balances by enforcing that, at any round t , the total amount of resource consumed remains close to the pacing target $b \cdot t$ with high probability.

Load balancing algorithm \mathcal{A}_x for any basis x

Without loss of generality, write $x = \{k, l\}$ with $c_{k,t_i} \geq c_{l,t_i}$. For any time period $t \geq t_i$, define $b_t^{\{k,l\}}$ as the total amount of resource consumed when selecting basis $\{k, l\}$ in the past $t - 1$ rounds. If basis $\{k, l\}$ is selected at time t , pull arm k if $b_t^{\{k,l\}} \leq n_{\{k,l\},t} \cdot b$ and pull arm l otherwise.

Observe that a basis $x = \{k, l\}$ is feasible for (3) if $\mu_k^c \geq b \geq \mu_l^c$. Moreover, the exploration and exploitation terms defined in Section 2 specialize to:

$$\text{obj}_{\{k,l\},t} = \left(\frac{\bar{c}_{k,t} - b}{\bar{c}_{k,t} - \bar{c}_{l,t}} \cdot \bar{r}_{l,t} + \frac{b - \bar{c}_{l,t}}{\bar{c}_{k,t} - \bar{c}_{l,t}} \cdot \bar{r}_{k,t} \right), \quad E_{\{k,l\},t} = \lambda \cdot \left(\frac{\bar{c}_{k,t} - b}{\bar{c}_{k,t} - \bar{c}_{l,t}} \cdot \epsilon_{l,t} + \frac{b - \bar{c}_{l,t}}{\bar{c}_{k,t} - \bar{c}_{l,t}} \cdot \epsilon_{k,t} \right),$$

provided that $\bar{c}_{k,t} \geq b \geq \bar{c}_{l,t}$, and we have:

$$\xi_{l,t}^{\{k,l\}} = \frac{\bar{c}_{k,t} - b}{\bar{c}_{k,t} - \bar{c}_{l,t}}, \xi_{k,t}^{\{k,l\}} = \frac{b - \bar{c}_{l,t}}{\bar{c}_{k,t} - \bar{c}_{l,t}}, \xi_l^{\{k,l\}} = \frac{\mu_k^c - b}{\mu_k^c - \mu_l^c}, \text{ and } \xi_k^{\{k,l\}} = \frac{b - \mu_l^c}{\mu_k^c - \mu_l^c}.$$

Regret Analysis. As stressed at the beginning of this section, UCB-Simplex may sometimes select an infeasible basis. We start by proving that this does not happen very often.

Lemma 13 *For any basis $x \notin \mathcal{B}$, we have: $\mathbb{E}[n_{x,T}] \leq \frac{\pi^2}{3\epsilon^2}$.*

Just like in Sections 3 and 4, a crucial property is that any suboptimal feasible basis is selected at most $O(\ln(T))$ times on average.

Lemma 14 *For any suboptimal basis $x \in \mathcal{B}$, we have:*

$$\mathbb{E}[n_{x,T}] \leq \left(\frac{4}{\epsilon}\right)^4 \cdot \frac{\ln(T)}{(\Delta_x)^2} + \frac{5\pi^2}{\epsilon^3}.$$

It remains to lower bound the expected total payoff derived when selecting any of the optimal bases. The major difficulty lies in the fact that the amounts of resource consumed, the rewards obtained, and the stopping time are correlated in a non-trivial way through the budget constraint and the decisions made in the past. This makes it difficult to study the expected total payoff derived when selecting optimal bases independently from the amounts of resource consumed and the rewards obtained when selecting suboptimal ones. However, a key point is that, by design, the decision made at Step 2 of UCB-Simplex is based solely on the past history associated with the basis selected at Step 1 because the load balancing algorithms are decoupled. For this reason, the analysis proceeds in two steps irrespective of the nature of the possibly many optimal bases. In a first step, we show that, for any basis x , the amount of resource consumed per round when selecting x stay close to the target b with high probability. This enables us to show that the ratios $(\mathbb{E}[n_{k,T}^x]/\mathbb{E}[n_{l,T}^x])_{k,l \in \text{supp}(x)}$ remain close to the optimal ones $(\xi_k^x/\xi_l^x)_{k,l \in \text{supp}(x)}$, as precisely stated below.

Lemma 15 *For any feasible basis $x = \{k, l\}$, and time period $t \geq t_i$, we have:*

$$\mathbb{P}[|b_t^x - n_{x,t} \cdot b| \geq u] \leq \frac{4}{\epsilon^2} \cdot \exp(-\epsilon^2 \cdot u) + \frac{2}{T^2}, \quad \forall u \geq 1,$$

which, in particular, implies that:

$$\begin{aligned} \mathbb{E}[n_{k,T}^x] &\geq \xi_k^x \cdot \mathbb{E}[n_{x,T}] - \frac{4}{\epsilon^5}, \\ \mathbb{E}[n_{l,T}^x] &\geq \xi_l^x \cdot \mathbb{E}[n_{x,T}] - \frac{4}{\epsilon^5}. \end{aligned} \tag{10}$$

In a second step, we show using Lemma 15 that, at the cost of an additive constant term in the regret bound, we may assume that the game last exactly T rounds. This enables us to combine Lemmas 13, 14, and 15 to establish the regret bound.

Proposition 16

$$R_{B,T} \leq \left(\frac{4}{\epsilon}\right)^4 \cdot \left(\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{1}{\Delta_x} \right) \cdot \ln(T) + \frac{2K}{\epsilon^3} \cdot \ln(T) + O(1).$$

Since the number of feasible bases for the linear program (3) is at most K^2 , we get the distribution-dependent regret bound $O(K^2 \cdot \frac{\ln(T)}{\Delta})$ where $\Delta = \min_{x \in \mathcal{B} \mid \Delta_x > 0} \Delta_x$. Along the same lines as in Sections 3 and 4, pushing the analysis further almost immediately implies a distribution-free regret bound.

Proposition 17

$$R_{B,T} \leq \left(\frac{4}{\epsilon}\right)^2 \cdot \sqrt{|\mathcal{B}| \cdot T \cdot \ln(T)}.$$

We conclude that $R_{B,T} = O(K \cdot \sqrt{T \cdot \ln(T)})$, where the hidden factors are independent of the underlying distributions $(\nu_k)_{k=1, \dots, K}$. Just like in Section 4, we stress that the dependence on K is not optimal since Agrawal and Devanur (2014) obtain a $\tilde{O}(\sqrt{K \cdot T})$ bound on regret, where the \tilde{O} notation hides factors logarithmic in T .

6. Extensions

6.1. A single limited resource

Deterministic amounts of resource consumption. If the amounts of resource consumed are deterministic, we can substitute the notation μ_k^c for c_k . Moreover, we can take $\lambda = 1$ and, going through the analysis of Lemma 5, we can slightly refine the bound as follows. Observe that, for any arm k such that $\Delta_k > 0$, $\mathbb{E}[n_{k,\tau^*}] \leq \frac{16}{(c_k)^2} \cdot \frac{\mathbb{E}[\ln(\tau^*)]}{(\Delta_k)^2} + \frac{\pi^2}{3}$. As a result, the regret bound derived in Proposition 6 turns into:

$$R_B \leq 16 \cdot \left(\sum_{k \mid \Delta_k > 0} \frac{1}{\Delta_k \cdot c_k} \right) \cdot \ln\left(\frac{B+1}{\epsilon}\right) + O(1),$$

which is identical (up to constant factors) to the bound obtained by Tran-Thanh et al. (2012).

Relaxing Assumption 2. If ϵ is unknown prior to starting the game, achieving a $O(\ln(B)^{1+\gamma})$ bound on regret for any $\gamma > 0$ is still possible by systematically adding an offset of $\ln(B)^{-\frac{\gamma}{4}}$ to the observed amounts of resource consumed and by taking $\lambda = 1 + \ln(B)^{\frac{\gamma}{4}}$. Proceeding this way, the analysis carried out in Lemma 5 is simplified because the amount of resource consumed at each step becomes almost surely no smaller than $\ln(B)^{-\frac{\gamma}{4}}$, thus making the disjunction $\bar{c}_{k,t} \leq \frac{\ln(B)^{-\frac{\gamma}{4}}}{2}$ unnecessary. Furthermore, observe that, for B large enough, we have:

$$\operatorname{argmax}_{k=1, \dots, K} \frac{\mu_k^r}{\mu_k^c + \ln(B)^{-\frac{\gamma}{4}}} \subset \operatorname{argmax}_{k=1, \dots, K} \frac{\mu_k^r}{\mu_k^c},$$

and

$$\frac{\mu_m^r}{\mu_m^c + \ln(B)^{-\frac{\gamma}{4}}} - \frac{\mu_l^r}{\mu_l^c + \ln(B)^{-\frac{\gamma}{4}}} > \Delta_l,$$

for any $m \in \operatorname{argmax}_{k=1, \dots, K} \frac{\mu_k^r}{\mu_k^c}$ and $l \notin \operatorname{argmax}_{k=1, \dots, K} \frac{\mu_k^r}{\mu_k^c}$. As a consequence, the bound derived in Lemma 5 turns into:

$$\mathbb{E}[n_{k,\tau^*}] \leq 4^4 \cdot \ln(B)^\gamma \cdot \frac{\mathbb{E}[\ln(\tau^*)]}{(\Delta_k)^2} + \frac{2\pi^2}{3},$$

for any k such that $\Delta_k > 0$. Going through the proof of Proposition 6, we obtain the asymptotic regret bound $R_B = O((\sum_{k \mid \Delta_k > 0} \frac{1}{\Delta_k}) \cdot \ln(B)^{1+\gamma})$.

6.2. Arbitrarily many limited resources whose consumption are deterministic

We propose another load balancing algorithm that couples bases together. This is key to get a better dependence on K because, otherwise, we have to study each basis independently from the other ones.

Load balancing algorithm \mathcal{A}_x for a feasible basis $x \in \mathcal{B}$

If basis x is selected at time t , pull any arm $a_t \in \operatorname{argmin}_{k \in \operatorname{supp}(x)} \frac{n_{k,t}}{\xi_k^x}$.

Observe that this load balancing algorithm runs in $O(K)$ computation time and requires $O(K)$ memory space. The shortcoming of this approach is that, if there happens to exist multiple optimal bases to (3), the optimal ratios for each optimal basis will not be preserved since we take account of the number of times we have pulled each arm k when selecting any other optimal basis (for which we strived to enforce different ratios). Hence, the following assumption will be required for the analysis.

Assumption 4 *There is a unique optimal basis to (3) x^* .*

Regret Analysis. We start by globally bounding, for each arm k , the number of times this arm can be pulled when selecting any of the suboptimal bases. This is in contrast to the analysis carried out in Section 4 where we bound the number of times each suboptimal basis has been selected. We will need the following notation:

$$\Delta_k = \min_{x \in \mathcal{B} \mid k \in \operatorname{supp}(x), x \neq x^*} \Delta_x,$$

for each arm k .

Lemma 18 *For any arm $k \notin x^*$, we have:*

$$\mathbb{E}[n_{k,\tau^*}] \leq \frac{16 \cdot r_{1,\dots,C}}{\epsilon^2} \cdot \frac{\mathbb{E}[\ln(\tau^*)]}{(\Delta_k)^2} + K \cdot \frac{\pi^2}{3}.$$

Lemma 19 *For any arm $k \in x^*$, we have:*

$$\mathbb{E}\left[\sum_{x \in \mathcal{B} \mid k \in x, x \neq x^*} n_{k,\tau^*}^x\right] \leq \frac{16 \cdot r_{1,\dots,C}}{\epsilon^2} \cdot \frac{\mathbb{E}[\ln(\tau^*)]}{(\Delta_k)^2} + K \cdot \frac{\pi^2}{3}.$$

In contrast to Section 4, we can only guarantee that the ratios $(n_{k,t}^x/n_{l,t}^x)_{k,l \in \operatorname{supp}(x)}$ remain close to the optimal ones $(\xi_k^x/\xi_l^x)_{k,l \in \operatorname{supp}(x)}$ at all times for the optimal basis $x = x^*$. This will not allow us to derive distribution-free regret bounds for this particular class of load balancing algorithms.

Lemma 20 *At any time t and for any arm $k \in \operatorname{supp}(x^*)$, we have:*

$$n_{k,t} \geq n_{x^*,t} \cdot \frac{\xi_k^{x^*}}{\sum_{l=1}^K \xi_l^{x^*}} - r_{1,\dots,C} \cdot \left(\sum_{x \in \mathcal{B}, x \neq x^*} n_{x,t} + 1 \right) \quad (11)$$

and

$$n_{k,t} \leq n_{x^*,t} \cdot \frac{\xi_k^{x^*}}{\sum_{l=1}^K \xi_l^{x^*}} + \sum_{x \in \mathcal{B}, x \neq x^*} n_{x,t} + 1. \quad (12)$$

Bringing everything together, we are now ready to establish regret bounds.

Proposition 21

$$R_{B(1), \dots, B(C)} \leq \frac{16 \cdot (r_1, \dots, C)^3}{\epsilon^3 \cdot b} \cdot \left(\sum_{k=1}^K \frac{1}{(\Delta_k)^2} \right) \cdot \ln\left(\frac{B+1}{\epsilon}\right) + O(1).$$

We derive a distribution-dependent regret bound of order $O(K \cdot \frac{\ln(B)}{(\Delta)^2})$ where $\Delta = \min_{x \in \mathcal{B} \mid \Delta_x > 0} \Delta_x$ but no non-trivial distribution-free regret bound.

7. Concluding remark

The existence of an algorithm with an expected bound on regret of order $O(K \cdot \ln(\min_i B(i)))$ in the case of multiple stochastic resources remains an open question which calls for the development of more efficient load balancing algorithms, in the same spirit as done in Section 6 in the case of multiple deterministic resources.

References

- S. Agrawal and N. R. Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the 15th ACM conference on Economics and Computation*, pages 989–1006, 2014.
- S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23, 2012.
- S. Agrawal, N. R. Devanur, L. Li, and N. Rangarajan. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Proceedings of the 29th Annual Conference on Learning Theory*, pages 4–18, 2016.
- K. Amin, S. Kale, G. Tesauro, and S. D. Turaga. Budgeted prediction with expert advice. In *29th AAAI Conference on Artificial Intelligence*, pages 2490–2496, 2015.
- P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a.
- P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.
- M. Babaioff, S. Dughmi, R. Kleinberg, and A. Slivkins. Dynamic pricing with limited supply. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 74–91, 2012.
- A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with knapsacks. In *54th IEEE Annual Symposium on Foundations of Computer Science*, pages 207–216, 2013.
- A. Badanidiyuru, J. Langford, and A. Slivkins. Resourceful contextual bandits. In *Proceedings of the 27th Annual Conference on Learning Theory*, volume 35, pages 1–26, 2014.
- D. Bertsimas and J. N. Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific, 1997.

- O. Besbes and A. Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.
- O. Besbes and A. Zeevi. Blind network revenue management. *Operations research*, 60(6):1537–1550, 2012.
- R. Combes, C. Jian, and R. Srikant. Bandits with budgets: Regret lower bounds and optimal algorithms. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 245–257, 2015.
- W. Ding, T. Qin, X. Zhang, and T. Liu. Multi-armed bandit with budget constraint and variable costs. In *27th AAAI Conference on Artificial Intelligence*, pages 232–238, 2013.
- M. Grötschel, L. Lovász, and A. Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2. Springer Science & Business Media, 2012.
- W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.
- K. Johnson, D. Simchi-Levi, and H. Wang. Online network revenue management using thompson sampling. *Working Paper*, 2015.
- R. Kleinberg and T. Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th IEEE Annual Symposium on Foundations of Computer Science*, pages 594–605, 2003.
- T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.
- C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999.
- H. Robbins. Some aspects of the sequential design of experiments. *Bulleting of the American Mathematical Society*, 58(5):527–535, 1952.
- A. Slivkins. Dynamic ad allocation: Bandits with budgets. *arXiv preprint arXiv:1306.0155*, 2013.
- L. Tran-Thanh, A. Chapman, A. Rogers, and N. R. Jennings. ϵ -first policies for budget-limited multi-armed bandits. In *24th AAAI Conference on Artificial Intelligence*, pages 1211–1216, 2010.
- L. Tran-Thanh, A. Chapman, A. Rogers, and N. R. Jennings. Knapsack based optimal policies for budget-limited multi-armed bandits. In *26th AAAI Conference on Artificial Intelligence*, pages 1134–1140, 2012.
- L. Tran-Thanh, C. Stavrogiannis, V. Naroditskiy, V. Robu, N. R. Jennings, and P. Key. Efficient regret bounds for online bid optimisation in budget-limited sponsored search auctions. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pages 809–818, 2014a.
- L. Tran-Thanh, S. Stein, A. Rogers, and N. R. Jennings. Efficient crowdsourcing of unknown experts using multi-armed bandits. *Artificial Intelligence*, 214:89–111, 2014b.

H. Wu, S. Srikant, X. Liu, and C. Jiang. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 433–441, 2015.

Y. Xia, W. Ding, X. Zhang, N. Yu, and T. Qin. Budgeted bandit problems with continuous random costs. In *Proceedings of the 7th Asian Conference on Machine Learning*, page 317332, 2015.

Appendix A. Proof of Lemma 2

The proof can be found in [Badanidiyuru et al. \(2013\)](#). For the sake of completeness, we reproduce it here. The optimization problem (3) is a linear program whose dual reads:

$$\begin{aligned}
 & \inf_{(\eta_i)_{i=1, \dots, C}} \quad \sum_{i=1}^C b(i) \cdot \eta_i \\
 & \text{subject to} \quad \sum_{i=1}^C \mu_k^c(i) \cdot \eta_i \geq \mu_k^r, \quad k = 1, \dots, K \\
 & \quad \quad \quad \eta_i \geq 0, \quad i = 1, \dots, C.
 \end{aligned} \tag{13}$$

Observe that (3) is feasible therefore (3) and (13) have the same optimal value. Note that (3) is bounded as $\xi_k \in [0, \frac{b(1)}{\mu_k^c(1)}]$ for any feasible point. Hence, (13) has an optimal solution $(\eta_1^*, \dots, \eta_C^*)$. Consider any non-anticipating algorithm. Let Z_t be the sum of the total payoff accumulated in rounds 1 to t plus the “cost” of the remaining resources, i.e. $Z_t = \sum_{\tau=1}^t r_{a_\tau, \tau} + \sum_{i=1}^C \eta_i^* \cdot (B(i) - \sum_{\tau=1}^t c_{a_\tau, \tau}(i))$. Observe that $(Z_t)_t$ is a supermartingale with respect to the filtration $(\mathcal{F}_t)_t$ as $\mathbb{E}[Z_t | \mathcal{F}_{t-1}] = \sum_{k=1}^K p_k^t \cdot (\mu_k^r - \sum_{i=1}^C \eta_i^* \cdot \mu_k^c(i)) + Z_{t-1} \leq Z_{t-1}$ where $p_k^t \in \mathcal{F}_{t-1}$ is determined by the algorithm and corresponds to the probability of pulling arm k at time t given the past. Moreover, $(Z_t)_t$ has bounded increments since $\mathbb{E}[|Z_t - Z_{t-1}| | \mathcal{F}_{t-1}] = \sum_{k=1}^K p_k^t \cdot \mathbb{E}[|r_{k,t} - \sum_{i=1}^C \eta_i^* \cdot c_{k,t}(i)|] \leq \sum_{k=1}^K p_k^t \cdot (1 + \sum_{i=1}^C \eta_i^*) = (1 + \sum_{i=1}^C \eta_i^*) < \infty$. We also have $\mathbb{E}[\tau^*] < \infty$ as:

$$\begin{aligned}
 \mathbb{E}[\tau^*] &= \sum_{t=1}^{\infty} \mathbb{P}[\tau^* \geq t] \\
 &\leq \sum_{t=1}^{\infty} \mathbb{P}[\sum_{\tau=1}^{t-1} c_{a_\tau, \tau}(i) \leq B(i), i = 1, \dots, C] \\
 &\leq 1 + \sum_{t=1}^{\infty} \mathbb{P}[\sum_{\tau=1}^t c_{a_\tau, \tau}(1) \leq t \cdot \min_k \mu_k^c(1) - (t \cdot \min_k \mu_k^c(1) - B(1))] \\
 &\leq (\frac{B(1)}{\min_k \mu_k^c(1)} + 2) + \sum_{t \geq \frac{B(1)}{\min_k \mu_k^c(1)}}^{\infty} \exp(-\frac{2(t \cdot \min_k \mu_k^c(1) - B(1))^2}{t}) \\
 &< \infty,
 \end{aligned}$$

where the third inequality results from an application of Lemma 1. By Doob's optional stopping theorem, $\mathbb{E}[Z_{\tau^*}] \leq \mathbb{E}[Z_0] = \sum_{i=1}^C \eta_i^* \cdot B(i)$. Observe that:

$$\begin{aligned} \mathbb{E}[Z_{\tau^*}] &= \mathbb{E}\left[\sum_{k=1}^K p_k^{\tau^*} \cdot (\mu_k^r - \sum_{i=1}^C \eta_i^* \cdot \mu_k^c(i)) + Z_{\tau^*-1}\right] \\ &\geq \mathbb{E}\left[-\sum_{i=1}^C \eta_i^* + \sum_{t=1}^{\tau^*-1} r_{a_t,t}\right]. \end{aligned}$$

Consider $l \in \operatorname{argmin}_{i=1,\dots,C} b(i)$ and define $\eta_l = \max_k \frac{\mu_k^r}{\mu_k^c(l)}$ and $\eta_i = 0$ for $i \neq l$. Since η is a feasible solution for (13), we have:

$$\begin{aligned} \sum_{i=1}^C \eta_i^* &= \frac{\sum_{i=1}^C b(i) \cdot \eta_i^*}{b(l)} \\ &\leq \frac{\sum_{i=1}^C b(i) \cdot \eta_i}{b(l)} \\ &\leq \max_k \frac{\mu_k^r}{\mu_k^c(l)}. \end{aligned}$$

We get $\mathbb{E}[Z_{\tau^*}] \geq \mathbb{E}[\sum_{t=1}^{\tau^*-1} r_{a_t,t}] - \max_{k,i} \frac{\mu_k^r}{\mu_k^c(i)}$ and finally:

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^{\tau^*-1} r_{a_t,t}\right] &\leq \sum_{i=1}^C \eta_i^* \cdot B(i) + \max_{k,i} \frac{\mu_k^r}{\mu_k^c(i)} \\ &\leq B \cdot \sum_{i=1}^C \eta_i^* \cdot b(i) + \frac{1}{\epsilon}. \end{aligned}$$

By strong duality, $\sum_{i=1}^C \eta_i^* \cdot b(i)$ is also the optimal value of (3).

Appendix B. Proof of Lemma 3

Observe that, for any resource $i \in \{1, \dots, C\}$, we have:

$$\begin{aligned} T \cdot \operatorname{obj}_{x^*} - \mathbb{E}\left[\sum_{t=1}^{\tau^*} r_{a_t,t}\right] &= (T - \mathbb{E}[\tau^*]) \cdot \operatorname{obj}_{x^*} \\ &\geq \mathbb{E}\left[\left(\sum_{t=\tau^*}^T c_{a_t,t}(i) + \sum_{t=1}^{\tau^*-1} c_{a_t,t}(i) - B(i)\right)_+\right] \cdot \operatorname{obj}_{x^*} \\ &= \mathbb{E}\left[\left(\sum_{t=1}^T \{c_{a_t,t}(i) - b(i)\}\right)_+\right] \cdot \operatorname{obj}_{x^*}, \end{aligned}$$

where the inequality is derived using $c_{a_t,t} \leq 1$ for all rounds t and $\sum_{t=1}^{\tau^*-1} c_{a_t,t}(i) \leq B(i)$. Hence, $(c_{a_t,t}(i))_{t \in \mathbb{N}}$ is an i.i.d. bounded stochastic process with mean $b(i)$, which implies that:

$$\mathbb{E}\left[\left(\sum_{t=1}^T \{c_{a_t,t}(i) - b(i)\}\right)_+\right] = \Omega(\sqrt{T}), \quad (14)$$

provided that $c_{a_t,t}(i)$ has positive variance, which is true if there exists at least one arm $k \in \text{supp}(x^*)$ such that $c_{k,t}(i)$ has positive variance or if there exist two arms $k, l \in \text{supp}(x^*)$ such that $c_{k,t}(i)$ and $c_{l,t}(i)$ are not almost surely equal to the same deterministic value. Strictly speaking, this is not enough to conclude that $R_{B(1), \dots, B(C), T} = \Omega(\sqrt{T})$ as $T \cdot \text{obj}_{x^*}$ is only an upper bound on the maximum achievable payoff. However, in Sections 3, 4, and 5, we show that, in fact, there exists an algorithm such that $\text{ER}_{\text{OPT}}(B(1), \dots, B(C), T) - \mathbb{E}[\sum_{t=1}^{\tau^*} r_{a_t,t}] = O(\ln(T))$ for all the cases considered in this paper, which, in combination with (14) and Lemma 2, implies that $R_{B(1), \dots, B(C), T} = \Omega(\sqrt{T})$.

Appendix C. Proof of Lemma 4

By definition of τ^* , we have $\sum_{t=1}^{\tau^*-1} c_{a_t,t} \leq B$. Taking expectations on both sides yields:

$$\begin{aligned}
 B &\geq \mathbb{E}\left[\sum_{t=1}^{\tau^*-1} c_{a_t,t}\right] \\
 &\geq \mathbb{E}\left[\sum_{t=1}^{\tau^*} c_{a_t,t}\right] - 1 \\
 &= \sum_{t=1}^{\infty} \mathbb{E}[I_{\tau^* \geq t} \cdot c_{a_t,t}] - 1 \\
 &= \sum_{t=1}^{\infty} \mathbb{E}[I_{\tau^* \geq t} \cdot \mathbb{E}[c_{a_t,t} \mid \mathcal{F}_{t-1}]] - 1 \\
 &= \sum_{t=1}^{\infty} \mathbb{E}[I_{\tau^* \geq t} \cdot \mu_{a_t}^c] - 1 \\
 &\geq \sum_{t=1}^{\infty} \mathbb{E}[I_{\tau^* \geq t} \cdot \epsilon] - 1 \\
 &= \mathbb{E}[\tau^*] \cdot \epsilon - 1,
 \end{aligned}$$

where we use the fact that $c_{k,t} \leq 1$ for all arms k to derive the second inequality, the fact that τ^* is a stopping time for the second equality, the fact that the algorithm is non-anticipating for the third equality, i.e. $a_t \in \mathcal{F}_{t-1}$, and Assumption 2 for the third inequality. We conclude that $\mathbb{E}[\tau^*] \leq \frac{B+1}{\epsilon}$.

Appendix D. Proof of Lemma 5

We break down the analysis in a series of facts. Consider any k such that $\Delta_k > 0$. We use the shorthand $\beta_k = 2 \cdot \frac{4^3}{\epsilon^4} \cdot (\frac{1}{\Delta_k})^2$.

Fact 1

$$\mathbb{E}[n_{k,\tau^*}] \leq 2\beta_k \cdot \mathbb{E}[\ln(\tau^*)] + \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{a_t=k} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}\right]. \quad (15)$$

Proof Define the random variable $T_k = \beta_k \cdot \ln(\tau^*)$. We have:

$$\begin{aligned}\mathbb{E}[n_{k,\tau^*}] &= \mathbb{E}[n_{k,\tau^*} \cdot I_{n_{k,\tau^*} < T_k}] + \mathbb{E}[n_{k,\tau^*} \cdot I_{n_{k,\tau^*} \geq T_k}] \\ &\leq \beta_k \cdot \mathbb{E}[\ln(\tau^*)] + \mathbb{E}[n_{k,\tau^*} \cdot I_{n_{k,\tau^*} \geq T_k}].\end{aligned}$$

Define T_k^* as the first time t such that $n_{k,t} \geq T_k$ and $T_k^* = \infty$ if no such t exists. We have:

$$\begin{aligned}\mathbb{E}[n_{k,\tau^*} \cdot I_{n_{k,\tau^*} \geq T_k}] &= \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{a_t=k} \cdot I_{n_{k,\tau^*} \geq T_k}\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T_k^*-1} I_{a_t=k} \cdot I_{n_{k,\tau^*} \geq T_k}\right] + \mathbb{E}\left[\sum_{t=T_k^*}^{\tau^*} I_{a_t=k} \cdot I_{n_{k,\tau^*} \geq T_k}\right] \\ &\leq \mathbb{E}[n_{k,T_k^*-1} \cdot I_{n_{k,\tau^*} \geq T_k}] + \mathbb{E}\left[\sum_{t=T_k^*}^{\tau^*} I_{a_t=k} \cdot I_{n_{k,t} \geq T_k}\right] \\ &\leq \beta_k \cdot \mathbb{E}[\ln(\tau^*)] + \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{a_t=k} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}\right],\end{aligned}$$

since, by definition of T_k^* , $n_{k,T_k^*-1} \leq T_k$ if T_k^* is finite, which is always true if $n_{k,\tau^*} \geq T_k$ (the sequence $(n_{k,t})_t$ is non-decreasing and τ^* is finite almost surely as a byproduct of Lemma 4). Conversely, $n_{k,t} \geq T_k \geq \beta_k \ln(t)$ for $t \in \{T_k^*, \dots, \tau^*\}$. Wrapping up, we obtain:

$$\mathbb{E}[n_{k,\tau^*}] \leq 2\beta_k \cdot \mathbb{E}[\ln(\tau^*)] + \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{a_t=k} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}\right].$$

■

Fact 1 enables us to assume that arm k has been pulled at least $\beta_k \ln(t)$ times out of the last t time periods. The remainder of this proof is dedicated to show that the second term of the right-hand side of (15) can be bounded by a constant. Let us first rewrite this term:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{a_t=k} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}\right] &\leq \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{k,t} + E_{k,t} \geq \text{obj}_{k^*,t} + E_{k^*,t}} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}\right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{k,t} \geq \text{obj}_k + E_{k,t}}\right] \tag{16}\end{aligned}$$

$$+ \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{k^*,t} \leq \text{obj}_{k^*} - E_{k^*,t}}\right] \tag{17}$$

$$+ \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{k^*} < \text{obj}_k + 2E_{k,t}} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}\right] \tag{18}$$

To derive this last inequality, simply observe that if $\text{obj}_{k,t} < \text{obj}_k + E_{k,t}$ and $\text{obj}_{k^*,t} > \text{obj}_{k^*} - E_{k^*,t}$ while $\text{obj}_{k,t} + E_{k,t} \geq \text{obj}_{k^*,t} + E_{k^*,t}$, it must be that $\text{obj}_{k^*} < \text{obj}_k + 2E_{k,t}$. Let us study (16), (17) and (18) separately.

Fact 2

$$\mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{k^*} < \text{obj}_k + 2E_{k,t}} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}\right] \leq \frac{2\pi^2}{3\epsilon^2}.$$

Proof We have:

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{k^*} < \text{obj}_k + 2E_{k,t}} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}\right] &\leq \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\bar{c}_{k,t} < \frac{\epsilon}{2}} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}\right] \\ &\quad + \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\bar{c}_{k,t} \geq \frac{\epsilon}{2}} \cdot I_{\text{obj}_{k^*} < \text{obj}_k + 2E_{k,t}} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}\right]. \end{aligned}$$

We upper bound the first term using the concentration inequalities of Lemma 1. First observe that:

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^{\infty} I_{\bar{c}_{k,t} < \frac{\epsilon}{2}} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}\right] &= \sum_{t=1}^{\infty} \mathbb{P}[\bar{c}_{k,t} < \frac{\epsilon}{2}; n_{k,t} \geq \beta_k \ln(t)] \\ &\leq \sum_{t=1}^{\infty} \sum_{s=\beta_k \ln(t)}^t \mathbb{P}[\bar{c}_{k,t} < \mu_k^c - (\mu_k^c - \frac{\epsilon}{2}); n_{k,t} = s] \\ &\leq \sum_{t=1}^{\infty} \sum_{s=\beta_k \ln(t)}^t \mathbb{P}[\bar{c}_{k,t} < \mu_k^c - \frac{\epsilon}{2}; n_{k,t} = s]. \end{aligned}$$

Denote by t_1, \dots, t_s the first s random times at which arm k is pulled (these random variables are finite almost surely). We have:

$$\mathbb{P}[\bar{c}_{k,t} < \mu_k^c - \frac{\epsilon}{2}; n_{k,t} = s] \leq \mathbb{P}\left[\sum_{l=1}^s c_{k,t_l} < s \cdot \mu_k^c - s \cdot \frac{\epsilon}{2}\right].$$

Observe that, for any $l \leq s$:

$$\begin{aligned} \mathbb{E}[c_{k,t_l} \mid c_{k,t_1}, \dots, c_{k,t_{l-1}}] &= \mathbb{E}\left[\sum_{\tau=1}^{\infty} I_{t_l=\tau} \cdot \mathbb{E}[c_{k,\tau} \mid \mathcal{F}_{\tau-1}] \mid c_{k,t_1}, \dots, c_{k,t_{l-1}}\right] \\ &= \mathbb{E}\left[\sum_{\tau=1}^{\infty} I_{t_l=\tau} \cdot \mu_k^c \mid c_{k,t_1}, \dots, c_{k,t_{l-1}}\right] \\ &= \mu_k^c \end{aligned}$$

since the algorithm is not randomized ($\{t_l = \tau\} \in \mathcal{F}_{\tau-1}$) and using the tower property. Hence, we can apply Lemma 1 to get:

$$\begin{aligned} \sum_{t=1}^{\infty} \mathbb{P}[\bar{c}_{k,t} < \frac{\epsilon}{2}; n_{k,t} \geq \beta_k \ln(t)] &\leq \sum_{t=1}^{\infty} \sum_{s=\beta_k \ln(t)}^{\infty} \exp(-s \cdot \frac{\epsilon^2}{2}) \\ &\leq \sum_{t=1}^{\infty} \frac{\exp(-\frac{\epsilon^2}{2} \beta_k \ln(t))}{1 - \exp(-\frac{\epsilon^2}{2})} \\ &\leq \frac{1}{1 - \exp(-\frac{\epsilon^2}{2})} \sum_{t=1}^{\infty} \frac{1}{t^2} \\ &\leq \frac{2\pi^2}{3\epsilon^2}, \end{aligned}$$

where we use the fact that $\beta_k \geq 2 \cdot \frac{4^3}{\epsilon^4} \cdot (\frac{\mu_k^c}{\mu_k^r})^2 \geq \frac{4}{\epsilon^2}$ for the third inequality and the fact that $\exp(-x) \leq 1 - \frac{x}{2}$ for $x \in [0, 1]$ for the last inequality.

As for the second term, observe that when both $n_{k,t} \geq \beta_k \ln(t)$ and $\bar{c}_{k,t} \geq \frac{\epsilon}{2}$, we have:

$$\begin{aligned} E_{k,t} &\leq (1 + \frac{1}{\epsilon}) \cdot \frac{2}{\epsilon} \cdot \sqrt{\frac{2}{\beta_k}} \\ &\leq \frac{\Delta_k}{2} \end{aligned}$$

since:

$$\begin{aligned} \beta_k &= 2 \cdot \frac{4^3}{\epsilon^4} \cdot (\frac{1}{\Delta_k})^2 \\ &\geq 2 \cdot (\frac{4}{\epsilon} \cdot (1 + \frac{1}{\epsilon}))^2 (\frac{1}{\Delta_k})^2. \end{aligned}$$

Hence, the second term is zero. ■

Let us now elaborate on (16).

Fact 3

$$\mathbb{E}[\sum_{t=1}^{\tau^*} I_{\text{obj}_{k,t} \geq \text{obj}_k + E_{k,t}}] \leq \frac{\pi^2}{3}.$$

Proof Note that if $\frac{\bar{r}_{k,t}}{\bar{c}_{k,t}} = \text{obj}_{k,t} \geq \text{obj}_k + E_{k,t} = \frac{\mu_k^r}{\mu_k^c} + E_{k,t}$, then either $\bar{r}_{k,t} \geq \mu_k^r + \epsilon_{k,t}$ or $\bar{c}_{k,t} \leq \mu_k^c - \epsilon_{k,t}$, otherwise we would have:

$$\begin{aligned} \frac{\bar{r}_{k,t}}{\bar{c}_{k,t}} - \frac{\mu_k^r}{\mu_k^c} &= \frac{(\bar{r}_{k,t} - \mu_k^r)\mu_k^c + (\mu_k^c - \bar{c}_{k,t})\mu_k^r}{\bar{c}_{k,t} \cdot \mu_k^c} \\ &< \frac{\epsilon_{k,t}}{\bar{c}_{k,t}} + \frac{\epsilon_{k,t}}{\bar{c}_{k,t} \cdot \epsilon} \\ &= E_{k,t}, \end{aligned}$$

a contradiction. Therefore:

$$\begin{aligned}
 \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{k,t} \geq \text{obj}_k + E_{k,t}}\right] &\leq \sum_{t=1}^{\infty} \mathbb{P}[\bar{r}_{k,t} \geq \mu_k^r + \epsilon_{k,t}] + \mathbb{P}[\bar{c}_{k,t} \leq \mu_k^c - \epsilon_{k,t}] \\
 &\leq \sum_{t=1}^{\infty} \sum_{s=1}^t \mathbb{P}[\bar{r}_{k,t} \geq \mu_k^r + \sqrt{\frac{2 \ln(t)}{s}} ; n_{k,t} = s] \\
 &\quad + \sum_{t=1}^{\infty} \sum_{s=1}^t \mathbb{P}[\bar{c}_{k,t} \leq \mu_k^c - \sqrt{\frac{2 \ln(t)}{s}} ; n_{k,t} = s] \\
 &= \sum_{t=1}^{\infty} \sum_{s=1}^t \mathbb{P}\left[\sum_{l=1}^s r_{k,t_l} \geq s \cdot \mu_k^r + \sqrt{s \cdot 2 \ln(t)} ; n_{k,t} = s\right] \\
 &\quad + \sum_{t=1}^{\infty} \sum_{s=1}^t \mathbb{P}\left[\sum_{l=1}^s c_{k,t_l} \leq s \cdot \mu_k^c - \sqrt{s \cdot 2 \ln(t)} ; n_{k,t} = s\right] \\
 &\leq \sum_{t=1}^{\infty} \sum_{s=1}^t 2 \exp(-4 \ln(t)) \\
 &= \frac{\pi^2}{3},
 \end{aligned}$$

where the random variables $(t_l)_l$ are defined similarly as in the proof of Fact 2 and the fourth inequality results from an application of Lemma 1. \blacksquare

It remains to bound (17).

Fact 4

$$\mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{k^*,t} \leq \text{obj}_{k^*} - E_{k^*,t}}\right] \leq \frac{\pi^2}{3}.$$

Proof We proceed along the same lines as in the proof of Fact 3. As a matter of fact, the situation is perfectly symmetric because, in the course of proving Fact 3, we do not rely on the fact that we have pulled arm k more than $\beta_k \ln(t)$ times at any time t . If $\frac{\bar{r}_{k^*,t}}{\bar{c}_{k^*,t}} = \text{obj}_{k^*,t} \leq \text{obj}_{k^*} - E_{k^*,t} = \frac{\mu_{k^*}^r}{\mu_{k^*}^c} - E_{k^*,t}$, then we have either $\bar{r}_{k^*,t} \leq \mu_{k^*}^r - \epsilon_{k^*,t}$ or $\bar{c}_{k^*,t} \geq \mu_{k^*}^c + \epsilon_{k^*,t}$, otherwise we would have:

$$\begin{aligned}
 \frac{\bar{r}_{k^*,t}}{\bar{c}_{k^*,t}} - \frac{\mu_{k^*}^r}{\mu_{k^*}^c} &= \frac{(\bar{r}_{k^*,t} - \mu_{k^*}^r) \mu_{k^*}^c + (\mu_{k^*}^c - \bar{c}_{k^*,t}) \mu_{k^*}^r}{\bar{c}_{k^*,t} \cdot \mu_{k^*}^c} \\
 &> -\frac{\epsilon_{k^*,t}}{\bar{c}_{k^*,t}} - \frac{\epsilon_{k^*,t}}{\bar{c}_{k^*,t} \cdot \lambda} \\
 &= -E_{k^*,t},
 \end{aligned}$$

a contradiction. Therefore:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{k^*,t} \leq \text{obj}_{k^*} - E_{k^*,t}}\right] &\leq \mathbb{E}\left[\sum_{t=1}^{\infty} I_{\bar{r}_{k^*,t} \leq \mu_{k^*}^r - \epsilon_{k,t}} + I_{\bar{c}_{k^*,t} \geq \mu_{k^*}^c + \epsilon_{k,t}}\right] \\
&\leq \sum_{t=1}^{\infty} \sum_{s=1}^t \mathbb{P}[\bar{r}_{k^*,t} \leq \mu_{k^*}^r - \sqrt{\frac{2 \ln(t)}{s}}; n_{k^*,t} = s] \\
&\quad + \sum_{t=1}^{\infty} \sum_{s=1}^t \mathbb{P}[\bar{c}_{k^*,t} \geq \mu_{k^*}^c + \sqrt{\frac{2 \ln(t)}{s}}; n_{k^*,t} = s] \\
&\leq \sum_{t=1}^{\infty} \sum_{s=1}^t \frac{2}{t^4} \\
&= \frac{\pi^2}{3},
\end{aligned}$$

where the third inequality is obtained using Lemma 1 in the same fashion as in Fact 3. ■

We conclude:

$$\mathbb{E}[n_{k,\tau^*}] \leq 2\beta_k \cdot \mathbb{E}[\ln(\tau^*)] + \frac{4\pi^2}{3\epsilon^2}.$$

Appendix E. Proof of Proposition 6

First observe that:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{\tau^*} r_{a_t,t}\right] &= \sum_{t=1}^{\infty} \mathbb{E}[I_{\tau^* \geq t} \cdot \mathbb{E}[r_{a_t,t} \mid \mathcal{F}_{t-1}]] \\
&= \sum_{t=1}^{\infty} \mathbb{E}[I_{\tau^* \geq t} \cdot \mu_{a_t}^r] \\
&= \sum_{t=1}^{\infty} \sum_{k=1}^K \mu_k^r \cdot \mathbb{E}[I_{\tau^* \geq t} \cdot I_{a_t=k}] \\
&= \sum_{k=1}^K \mu_k^r \cdot \mathbb{E}\left[\sum_{t=1}^{\infty} I_{\tau^* \geq t} \cdot I_{a_t=k}\right] \\
&= \sum_{k=1}^K \mu_k^r \cdot \mathbb{E}[n_{k,\tau^*}],
\end{aligned}$$

since τ^* is a stopping time. Plugging this equality into (7) yields:

$$\begin{aligned}
R_B &\leq B \cdot \frac{\mu_{k^*}^r}{\mu_{k^*}^c} - \sum_{k=1}^K \mu_k^r \cdot \mathbb{E}[n_{k,\tau^*}] + O(1) \\
&\leq \frac{\mu_{k^*}^r}{\mu_{k^*}^c} \cdot (B - \sum_{k \mid \Delta_k=0} \mu_k^c \cdot \mathbb{E}[n_{k,\tau^*}]) - \sum_{k \mid \Delta_k>0} \mu_k^r \cdot \mathbb{E}[n_{k,\tau^*}] + O(1).
\end{aligned}$$

Along the same lines as for the rewards, we can show that $\mathbb{E}[\sum_{t=1}^{\tau^*} c_{a_t, t}] = \sum_{k=1}^K \mu_k^c \cdot \mathbb{E}[n_{k, \tau^*}]$. By definition of τ^* , we have $B \leq \sum_{t=1}^{\tau^*} c_{a_t, t}$ almost surely. Taking expectations on both sides yields:

$$\begin{aligned} B &\leq \sum_{k=1}^K \mu_k^c \cdot \mathbb{E}[n_{k, \tau^*}] \\ &\leq \sum_{k \mid \Delta_k=0} \mu_k^c \cdot \mathbb{E}[n_{k, \tau^*}] + \sum_{k \mid \Delta_k>0} \mu_k^c \cdot \mathbb{E}[n_{k, \tau^*}]. \end{aligned}$$

Plugging this inequality back into the regret bound, we get:

$$\begin{aligned} R_B &\leq \sum_{k \mid \Delta_k>0} \left(\frac{\mu_{k^*}^r}{\mu_{k^*}^c} \cdot \mu_k^c - \mu_k^r \right) \cdot \mathbb{E}[n_{k, \tau^*}] + O(1) \\ &\leq \sum_{k \mid \Delta_k>0} \mu_k^c \cdot \Delta_k \cdot \mathbb{E}[n_{k, \tau^*}] + O(1) \\ &\leq \sum_{k \mid \Delta_k>0} \Delta_k \cdot \mathbb{E}[n_{k, \tau^*}] + O(1), \end{aligned} \tag{19}$$

since $\mu_k^c \leq 1$ for all arms k . Using the upper bound of Lemma 4, the concavity of the logarithmic function, and Lemma 5, we derive:

$$\begin{aligned} R_B &\leq \left(\frac{4}{\epsilon}\right)^4 \cdot \left(\sum_{k \mid \Delta_k>0} \frac{1}{\Delta_k} \right) \cdot \ln\left(\frac{B+1}{\epsilon}\right) + \frac{4\pi^2}{3\epsilon^2} \cdot \left(\sum_{k \mid \Delta_k>0} \Delta_k \right) + O(1) \\ &\leq \left(\frac{4}{\epsilon}\right)^4 \cdot \left(\sum_{k \mid \Delta_k>0} \frac{1}{\Delta_k} \right) \cdot \ln\left(\frac{B+1}{\epsilon}\right) + K \cdot \frac{4\pi^2}{3\epsilon^3} + O(1), \end{aligned} \tag{20}$$

since $\Delta_k \leq \frac{\mu_{k^*}^r}{\mu_{k^*}^c} \leq \frac{1}{\epsilon}$ for any arm k .

Appendix F. Proof of Proposition 7

To get the distribution-free bound, we start from inequality (19) (derived in the proof of Proposition 6) and apply Lemma 5 only if Δ_k is big enough, taking into account that:

$$\sum_{k=1}^K \mathbb{E}[n_{k, \tau^*}] = \mathbb{E}[\tau^*] \leq \frac{B+1}{\epsilon}.$$

Specifically, we have:

$$\begin{aligned}
R_B &\leq \sup_{\substack{(n_1, \dots, n_K) \geq 0 \\ \sum_{k=1}^K n_k \leq \frac{B+1}{\epsilon}}} \left\{ \sum_{k \mid \Delta_k > 0} \min(\Delta_k \cdot n_k, \left(\frac{4}{\epsilon}\right)^4 \cdot \frac{\ln(\frac{B+1}{\epsilon})}{\Delta_k} + \frac{4\pi^2}{3\epsilon^2} \cdot \Delta_k) \right\} + O(1) \\
&\leq \sup_{\substack{(n_1, \dots, n_K) \geq 0 \\ \sum_{k=1}^K n_k \leq \frac{B+1}{\epsilon}}} \left\{ \sum_{k \mid \Delta_k > 0} \min(\Delta_k \cdot n_k, \left(\frac{4}{\epsilon}\right)^4 \cdot \frac{\ln(\frac{B+1}{\epsilon})}{\Delta_k}) \right\} + K \cdot \frac{4\pi^2}{3\epsilon^3} + O(1) \\
&\leq \sup_{\substack{(n_1, \dots, n_K) \geq 0 \\ \sum_{k=1}^K n_k \leq \frac{B+1}{\epsilon}}} \left\{ \sum_{k \mid \Delta_k > 0} \left(\frac{4}{\epsilon}\right)^2 \cdot \sqrt{n_k \cdot \ln(\frac{B+1}{\epsilon})} \right\} + O(1) \\
&= \left(\frac{4}{\epsilon}\right)^2 \cdot \sqrt{\ln(\frac{B+1}{\epsilon})} \cdot \sup_{\substack{(n_1, \dots, n_K) \geq 0 \\ \sum_{k=1}^K n_k \leq \frac{B+1}{\epsilon}}} \left\{ \sum_{k=1}^K \sqrt{n_k} \right\} + O(1) \\
&\leq \left(\frac{4}{\epsilon}\right)^2 \cdot \sqrt{K \cdot \frac{B+1}{\epsilon} \cdot \ln(\frac{B+1}{\epsilon})} + O(1),
\end{aligned}$$

where the third inequality is derived by maximizing on $\Delta_k \geq 0$ for all arms k and the last inequality results from an application of the Cauchy Schwarz inequality.

Appendix G. Proof of Lemma 8

Consider a basis $x \in \mathcal{B}$ and a time period t . For \mathcal{A}_x to be well-defined, we need to show that there always exists an arm $k \in \text{supp}(x)$ such that $n_{k,t}^x \leq n_{x,t} \cdot \frac{\xi_k^x}{\sum_{l=1}^K \xi_l^x}$. Suppose there is none, we have:

$$\begin{aligned}
n_{x,t} &= \sum_{k \in \text{supp}(x)} n_{k,t}^x \\
&> \sum_{k \in \text{supp}(x)} n_{x,t} \cdot \frac{\xi_k^x}{\sum_{l=1}^K \xi_l^x} \\
&= n_{x,t} \cdot \sum_{k \in \text{supp}(x)} \frac{\xi_k^x}{\sum_{l=1}^K \xi_l^x} \\
&= n_{x,t},
\end{aligned}$$

a contradiction. Moreover, we have, at any time t and for any arm $k \in \text{supp}(x)$:

$$n_{k,t}^x \leq n_{x,t} \cdot \frac{\xi_k^x}{\sum_{l=1}^K \xi_l^x} + 1.$$

Indeed, suppose otherwise and define $t^* \leq t$ as the last time arm k was pulled. Since $(n_{x,\tau})_{\tau=1,\dots,t}$ is a non-decreasing sequence, we have:

$$\begin{aligned} n_{k,t^*}^x &= n_{k,t}^x - 1 \\ &> n_{x,t} \cdot \frac{\xi_k^x}{\sum_{l=1}^K \xi_l^x} \\ &\geq n_{x,t^*} \cdot \frac{\xi_k^x}{\sum_{l=1}^K \xi_l^x}, \end{aligned}$$

which shows by definition that arm k could not have been pulled at time t^* . We also derive as a byproduct that, at any time t and for any arm $k \in \text{supp}(x)$:

$$n_{x,t} \cdot \frac{\xi_k^x}{\sum_{l=1}^K \xi_l^x} - r_{1,\dots,C} \leq n_{k,t}^x,$$

since $n_{x,t} = \sum_{k \in \text{supp}(x)} n_{k,t}^x$ and since a basis involves at most $r_{1,\dots,C}$ arms.

Appendix H. Proof of Lemma 9

By definition of τ^* , we have $\sum_{t=1}^{\tau^*-1} c_{a_t,t}(1) \leq B$. Using Assumption 2 and the fact that $(c_{k,t}(1))_{t=1,\dots,T}$ are deterministic values larger than ϵ for all arms k , we get $(\tau^* - 1) \cdot \epsilon \leq B$. Taking expectations on both sides yields:

$$\mathbb{E}[\tau^*] \leq \frac{B+1}{\epsilon}.$$

Appendix I. Proof of Lemma 10

Consider any suboptimal basis $x \in \mathcal{B}$. The proof is along the same lines as for Lemma 5 and follows the exact same steps. We use the shorthand $\beta_x = \frac{8 \cdot r_{1,\dots,C}}{\epsilon^2} \cdot (\frac{1}{\Delta_x})^2$.

Fact 5

$$\mathbb{E}[n_{x,\tau^*}] \leq 2\beta_x \cdot \mathbb{E}[\ln(\tau^*)] + \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{x_t=x} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}\right]. \quad (21)$$

We omit the proof as it is analogous to the proof of Fact 1. As in Lemma 5, we break down the second term in the right-hand side in three terms and bound each of them by a constant:

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{x_t=x} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}\right] &\leq \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x,t} + E_{x,t} \geq \text{obj}_{x^*,t} + E_{x^*,t}} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}\right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}}\right] \end{aligned} \quad (22)$$

$$+ \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x^*,t} \leq \text{obj}_{x^*} - E_{x^*,t}}\right] \quad (23)$$

$$+ \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x^*} < \text{obj}_x + 2E_{x,t}} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}\right]. \quad (24)$$

Fact 6

$$\mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x^*} < \text{obj}_x + 2E_{x,t}} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}\right] = 0.$$

Proof If $\text{obj}_{x^*} < \text{obj}_x + 2E_{x,t}$, we get:

$$\begin{aligned} \frac{\Delta_x}{2} &< \sum_{k \in \text{supp}(x)} \xi_k^x \cdot \sqrt{\frac{2 \ln(t)}{n_{k,t}}} \\ &\leq \sum_{k \in \text{supp}(x)} \xi_k^x \cdot \sqrt{\frac{2 \ln(t)}{r_{1,\dots,C} + n_{k,t}^x}} \\ &\leq \sqrt{\sum_{k \in \text{supp}(x)} \xi_k^x} \cdot \sum_{k \in \text{supp}(x)} \sqrt{\xi_k^x} \cdot \sqrt{\frac{2 \ln(t)}{n_{x,t}}}, \end{aligned}$$

where we use (8) and Lemma 8. This implies:

$$n_{x,t} < 8 \cdot r_{1,\dots,C} \cdot \left(\frac{\sum_{k \in \text{supp}(x)} \xi_k^x}{\Delta_x} \right)^2 \cdot \ln(t),$$

using the Cauchy–Schwarz inequality and the fact that a basis involves at most $r_{1,\dots,C}$ arms. Note that this inequality also holds if $x = \{k\}$. Now observe that:

$$\begin{aligned} \sum_{k \in \text{supp}(x)} \xi_k^x &\leq \min_i \sum_{k \in \text{supp}(x)} \frac{c_k(i)}{\epsilon} \cdot \xi_k^x \\ &\leq \frac{\min_i b(i)}{\epsilon} \\ &\leq \frac{1}{\epsilon}. \end{aligned}$$

We obtain:

$$\begin{aligned} n_{x,t} &< 8 \cdot \frac{r_{1,\dots,C}}{\epsilon^2} \cdot \frac{1}{(\Delta_x)^2} \cdot \ln(t) \\ &= \beta_x \cdot \ln(t). \end{aligned}$$

■

Fact 7

$$\mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}}\right] \leq r_{1,\dots,C} \cdot \frac{\pi^2}{6}.$$

Proof If $\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}$, there must exist $k \in \text{supp}(x)$ such that $\bar{r}_{k,t} \geq \mu_k^r + \epsilon_{k,t}$, otherwise:

$$\begin{aligned} \text{obj}_{x,t} - \text{obj}_x &= \sum_{k \in \text{supp}(x)} (\bar{r}_{k,t} - \mu_k^r) \cdot \xi_k^x \\ &< \sum_{k \in \text{supp}(x)} \epsilon_{k,t} \cdot \xi_k^x \\ &= E_{x,t}. \end{aligned}$$

We obtain:

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}}\right] &\leq \sum_{k \in \text{supp}(x)} \sum_{t=1}^{\infty} \mathbb{P}[\bar{r}_{k,t} \geq \mu_k^r + \epsilon_{k,t}] \\ &\leq r_1, \dots, C \cdot \frac{\pi^2}{6}, \end{aligned}$$

where the last inequality is derived along the same lines as in the proof of Fact 3. ■

Fact 8

$$\mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x^*,t} \leq \text{obj}_{x^*} - E_{x^*,t}}\right] \leq r_1, \dots, C \cdot \frac{\pi^2}{6}.$$

Proof Similar to Fact 7. ■

Appendix J. Proof of Proposition 11

The proof proceeds along the same lines as for Proposition 6. We start by refining (4):

$$\begin{aligned} R_{B(1), \dots, B(C)} &\leq B \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \mathbb{E}\left[\sum_{t=1}^{\tau^*} r_{a_t,t}\right] + O(1) \\ &= B \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \sum_{t=1}^{\infty} \mathbb{E}[I_{\tau^* \geq t} \cdot \sum_{k=1}^K \sum_{x \in \mathcal{B}} r_{k,t} \cdot I_{x_t=x, a_t=k}] + O(1) \\ &= B \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \sum_{t=1}^{\infty} \mathbb{E}[I_{\tau^* \geq t} \cdot \sum_{k=1}^K \sum_{x \in \mathcal{B}} I_{x_t=x, a_t=k} \cdot \mathbb{E}[r_{k,t} \mid \mathcal{F}_{t-1}]] + O(1) \\ &= B \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \sum_{x \in \mathcal{B}} \sum_{k=1}^K \mu_k^r \cdot \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{x_t=x, a_t=k}\right] + O(1) \\ &= B \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \sum_{x \in \mathcal{B}} \sum_{k=1}^K \mu_k^r \cdot \mathbb{E}[n_{k,\tau^*}^x] + O(1), \end{aligned}$$

where we use the fact that x_t and a_t are determined by the events of the first $t - 1$ rounds and that τ^* is a stopping time. Using the properties of the load balancing algorithms established in Lemma

8, we derive:

$$\begin{aligned}
R_{B(1), \dots, B(C)} &\leq B \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \sum_{x \in \mathcal{B}} \sum_{k \in \text{supp}(x)} \left\{ \mu_k^r \cdot \frac{\xi_k^x}{\sum_{l \in \text{supp}(x)} \xi_l^x} \cdot \mathbb{E}[n_{x, \tau^*}] - r_{1, \dots, C} \right\} + O(1) \\
&= B \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \sum_{x \in \mathcal{B}} \left\{ \frac{\mathbb{E}[n_{x, \tau^*}]}{\sum_{l \in \text{supp}(x)} \xi_l^x} \cdot \left(\sum_{k \in \text{supp}(x)} \mu_k^r \cdot \xi_k^x \right) - (r_{1, \dots, C})^2 \right\} + O(1) \\
&= \left(\sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} \right) \cdot \left(B - \sum_{x \in \mathcal{B} \mid \Delta_x = 0} \frac{\mathbb{E}[n_{x, \tau^*}]}{\sum_{l \in \text{supp}(x)} \xi_l^x} \right) \\
&\quad - \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \left\{ \left(\sum_{k \in \text{supp}(x)} \mu_k^r \cdot \xi_k^x \right) \cdot \frac{\mathbb{E}[n_{x, \tau^*}]}{\sum_{l \in \text{supp}(x)} \xi_l^x} \right\} + O(1).
\end{aligned}$$

Now observe that, by definition, at least one resource is exhausted at time τ^* . Hence, there exists $i \in \{1, \dots, C\}$ such that the following holds almost surely:

$$\begin{aligned}
B(i) &\leq \sum_{x \in \mathcal{B}} \sum_{k \in \text{supp}(x)} c_k(i) \cdot n_{k, \tau^*}^x \\
&\leq \sum_{x \in \mathcal{B}} \sum_{k \in \text{supp}(x)} \left[c_k(i) \cdot \left(\frac{\xi_k^x}{\sum_{l \in \text{supp}(x)} \xi_l^x} \cdot n_{x, \tau^*} + 1 \right) \right] \\
&= |\mathcal{B}| \cdot r_{1, \dots, C} + \sum_{x \in \mathcal{B}} \frac{n_{x, \tau^*}}{\sum_{l \in \text{supp}(x)} \xi_l^x} \cdot \sum_{k \in \text{supp}(x)} c_k(i) \cdot \xi_k^x \\
&\leq |\mathcal{B}| \cdot r_{1, \dots, C} + b(i) \cdot \sum_{x \in \mathcal{B}} \frac{n_{x, \tau^*}}{\sum_{l \in \text{supp}(x)} \xi_l^x},
\end{aligned}$$

where we use Lemma 8 again, the fact that any basis $x \in \mathcal{B}$ satisfies all the constraints of (3), and Assumption 2. We conclude that the inequality:

$$\sum_{x \in \mathcal{B} \mid \Delta_x = 0} \frac{n_{x, \tau^*}}{\sum_{l \in \text{supp}(x)} \xi_l^x} \geq B - \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{n_{x, \tau^*}}{\sum_{l \in \text{supp}(x)} \xi_l^x} - \frac{|\mathcal{B}| \cdot r_{1, \dots, C}}{b}$$

holds almost surely. Taking expectations on both sides and plugging the result back into the regret bound yields:

$$R_{B(1), \dots, B(C)} \leq \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{(\sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \sum_{k=1}^K \mu_k^r \cdot \xi_k^x)}{\sum_{l \in \text{supp}(x)} \xi_l^x} \cdot \mathbb{E}[n_{x, \tau^*}] \quad (25)$$

$$\begin{aligned}
&+ \left(\sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} \right) \cdot \frac{|\mathcal{B}| \cdot r_{1, \dots, C}}{b} + O(1) \\
&\leq \frac{1}{b} \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \Delta_x \cdot \mathbb{E}[n_{x, \tau^*}] + O(1), \quad (26)
\end{aligned}$$

where we use the fact that:

$$\begin{aligned} \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} &\leq \min_{i=1, \dots, C} \sum_{k=1}^K \frac{c_k(i)}{\epsilon} \cdot \xi_k^{x^*} \\ &\leq \min_{i=1, \dots, C} \frac{b(i)}{\epsilon} \\ &\leq \frac{1}{\epsilon}, \end{aligned}$$

and that, for any basis $x \in \mathcal{B}$, at least one of the first C inequalities is binding in (3), which implies that there exists $i \in \{1, \dots, C\}$ such that:

$$\begin{aligned} \sum_{k=1}^K \xi_k^x &\geq \sum_{k=1}^K c_k(i) \cdot \xi_k^x \\ &= b(i) \\ &\geq b. \end{aligned}$$

Using Lemma 10, Lemma 9, and the concavity of the logarithmic function, we obtain:

$$\begin{aligned} R_{B(1), \dots, B(C)} &\leq \frac{16 \cdot r_{1, \dots, C}}{b \cdot \epsilon^2} \cdot \left(\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{1}{\Delta_x} \right) \cdot \mathbb{E}[\ln(\tau^*)] + \frac{\pi^2 \cdot r_{1, \dots, C}}{3 \cdot b} \cdot \left(\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \Delta_x \right) + O(1) \\ &\leq \frac{16 \cdot r_{1, \dots, C}}{b \cdot \epsilon^2} \cdot \left(\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{1}{\Delta_x} \right) \cdot \ln\left(\frac{B+1}{\epsilon}\right) + O(1). \end{aligned}$$

Appendix K. Proof of Proposition 12

Along the same lines as for the case of a single limited resource, we start from inequality (26) derived in the proof of Proposition 11 and apply Lemma 10 only if Δ_x is big enough taking into account the fact that:

$$\sum_{x \in \mathcal{B}} \mathbb{E}[n_{x, \tau^*}] \leq \mathbb{E}[\tau^*] \leq \frac{B+1}{\epsilon}.$$

Specifically, we have:

$$\begin{aligned}
R_B &\leq \frac{1}{b} \cdot \sup_{\substack{(n_x)_{x \in \mathcal{B}} \geq 0 \\ \sum_{x \in \mathcal{B}} n_x \leq \frac{B+1}{\epsilon}}} \left\{ \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \min\left(\Delta_x \cdot n_x, \frac{16 \cdot r_{1,\dots,C}}{\epsilon^2} \cdot \frac{\ln(\frac{B+1}{\epsilon})}{\Delta_x} + \frac{\pi^2 \cdot r_{1,\dots,C}}{3} \cdot \Delta_x\right) \right\} + O(1) \\
&\leq \frac{1}{b} \cdot \sup_{\substack{(n_x)_{x \in \mathcal{B}} \geq 0 \\ \sum_{x \in \mathcal{B}} n_x \leq \frac{B+1}{\epsilon}}} \left\{ \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \min\left(\Delta_x \cdot n_x, \frac{16 \cdot r_{1,\dots,C}}{\epsilon^2} \cdot \frac{\ln(\frac{B+1}{\epsilon})}{\Delta_x}\right) \right\} + |\mathcal{B}| \cdot \frac{\pi^2 \cdot r_{1,\dots,C}}{3 \cdot \epsilon} + O(1) \\
&\leq \frac{1}{b} \cdot \sup_{\substack{(n_x)_{x \in \mathcal{B}} \geq 0 \\ \sum_{x \in \mathcal{B}} n_x \leq \frac{B+1}{\epsilon}}} \left\{ \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \sqrt{\frac{16 \cdot r_{1,\dots,C}}{\epsilon^2} \cdot \ln\left(\frac{B+1}{\epsilon}\right) \cdot n_x} \right\} + O(1) \\
&\leq \frac{1}{b} \cdot \frac{4 \cdot \sqrt{r_{1,\dots,C}}}{\epsilon} \cdot \sqrt{\ln\left(\frac{B+1}{\epsilon}\right)} \cdot \sup_{\substack{(n_x)_{x \in \mathcal{B}} \geq 0 \\ \sum_{x \in \mathcal{B}} n_x \leq \frac{B+1}{\epsilon}}} \left\{ \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \sqrt{n_x} \right\} + O(1) \\
&\leq \frac{4}{b \cdot \epsilon} \cdot \sqrt{r_{1,\dots,C} \cdot |\mathcal{B}| \cdot \frac{B+1}{\epsilon} \cdot \ln\left(\frac{B+1}{\epsilon}\right)},
\end{aligned}$$

where we maximize over each $\Delta_x \geq 0$ to derive the third inequality and we use Cauchy-Schwartz for the last inequality.

Appendix L. Proof of Lemma 13

Consider $x \notin \mathcal{B}$. Without loss of generality, we can assume that $x = \{k, l\}$ and $\mu_k^c, \mu_l^c > b$ (the situation is symmetric if the reverse inequality holds). If x is selected at time t , either $\bar{c}_{k,t} \leq b$ or $\bar{c}_{l,t} \leq b$, otherwise x would have been infeasible for (6). Thus, using (9):

$$\begin{aligned}
\mathbb{E}[n_{x,T}] &\leq \mathbb{E}\left[\sum_{t=t_i}^T I_{x_t=x} \cdot I_{n_{k,t} \geq \frac{1}{\epsilon^2} \ln(t)} \cdot I_{n_{l,t} \geq \frac{1}{\epsilon^2} \ln(t)}\right] \\
&\leq \sum_{t=t_i}^{\infty} \mathbb{P}[\bar{c}_{k,t} \leq b, n_{k,t} \geq \frac{1}{\epsilon^2} \ln(t)] + \mathbb{P}[\bar{c}_{l,t} \leq b, n_{l,t} \geq \frac{1}{\epsilon^2} \ln(t)].
\end{aligned}$$

Following the same recipe as in the proof of Fact 2, we conclude:

$$\mathbb{E}[n_{x,T}] \leq \frac{\pi^2}{3\epsilon^2}.$$

Appendix M. Proof of Lemma 14

Consider any suboptimal basis $x \in \mathcal{B}$. The proof is along the same lines as for Lemmas 5 and 10. We break down the analysis in a series of facts where we emphasize the main differences. We start off with an inequality analogous to Fact 1. The only difference lies in the initialization step which essentially guarantees that x^* is feasible for (6) with high probability. We use the shorthand $\beta_x = \frac{1}{2} \cdot \left(\frac{4}{\epsilon}\right)^4 \cdot \left(\frac{1}{\Delta_x}\right)^2$.

Fact 9

$$\begin{aligned} \mathbb{E}[n_{x,\tau^*}] &\leq 2\beta_x \cdot \ln(T) + \frac{\pi^2}{3\epsilon^2} \\ &\quad + \mathbb{E}\left[\sum_{t=t_i}^T I_{x_t=x} \cdot I_{n_{x,t} \geq \beta_x \ln(t)} \cdot I_{x^* \in \mathcal{B}_t}\right]. \end{aligned} \quad (27)$$

Proof For any suboptimal feasible basis x , define $T_x = \beta_x \cdot \ln(T)$. Without loss of generality, we can write $x^* = \{k^*, l^*\}$ with $\mu_{k^*}^c > b > \mu_{l^*}^c$. We start along same lines as in Fact 1 of Lemma 5 substituting k for x and using (9) to get:

$$\mathbb{E}[n_{x,T}] \leq 2\beta_x \cdot \ln(T) + \mathbb{E}\left[\sum_{t=t_i}^T I_{x_t=x} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}\right],$$

This further yields:

$$\begin{aligned} \mathbb{E}[n_{x,T}] &\leq 2\beta_x \cdot \ln(T) + \mathbb{E}\left[\sum_{t=t_i}^T I_{c_{k^*},t \leq b}\right] + \mathbb{E}\left[\sum_{t=t_i}^T I_{c_{l^*},t \geq b}\right] \\ &\quad + \mathbb{E}\left[\sum_{t=t_i}^T I_{x_t=x} \cdot I_{n_{x,t} \geq \beta_x \ln(t)} \cdot I_{x^* \in \mathcal{B}_t}\right]. \end{aligned}$$

We bound the second and third terms appearing in the right-hand side along the same lines as in Lemma 13 using the fact that $n_{k^*,t}, n_{l^*,t} \geq \frac{1}{\epsilon^2} \ln(t)$ as a result of the initialization step. ■

The remainder of this proof is dedicated to show that the last term in (27) can be bounded by a constant. This term can be broken down in three terms similarly as in Lemmas 5 and 10.

$$\begin{aligned} \mathbb{E}\left[\sum_{t=t_i}^T I_{x_t=x} \cdot I_{n_{x,t} \geq \beta_x \ln(t)} \cdot I_{x^* \in \mathcal{B}_t}\right] &\leq \mathbb{E}\left[\sum_{t=t_i}^T I_{\text{obj}_{x,t} + E_{x,t} \geq \text{obj}_{x^*,t} + E_{x^*,t}} \cdot I_{n_{x,t} \geq \beta_x \ln(t)} \cdot I_{x \in \mathcal{B}_t, x^* \in \mathcal{B}_t}\right] \\ &\leq \mathbb{E}\left[\sum_{t=t_i}^T I_{\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}} \cdot I_{x \in \mathcal{B}_t}\right] \end{aligned} \quad (28)$$

$$+ \mathbb{E}\left[\sum_{t=t_i}^T I_{\text{obj}_{x^*,t} \leq \text{obj}_{x^*} - E_{x^*,t}} \cdot I_{x^* \in \mathcal{B}_t}\right] \quad (29)$$

$$+ \mathbb{E}\left[\sum_{t=t_i}^T I_{\text{obj}_{x^*} < \text{obj}_x + 2E_{x,t}} \cdot I_{x \in \mathcal{B}_t} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}\right]. \quad (30)$$

We carefully study each term separately.

Fact 10

$$\mathbb{E}\left[\sum_{t=t_i}^T I_{\text{obj}_{x^*} < \text{obj}_x + 2E_{x,t}} \cdot I_{x \in \mathcal{B}_t} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}\right] \leq 2 + \frac{\pi^2}{\epsilon^3}.$$

Proof Without loss of generality, we can write $x = \{k, l\}$ with $\mu_k^c > b > \mu_l^c$. Using the shorthand $\alpha_x = 8 \cdot (\frac{\lambda}{\Delta_x})^2$, we have:

$$\begin{aligned}
& \mathbb{E} \left[\sum_{t=t_i}^T I_{\text{obj}_{x^*} < \text{obj}_x + 2E_{x,t}} \cdot I_{x \in \mathcal{B}_t} \cdot I_{n_{x,t} \geq \beta_x \cdot \ln(t)} \right] \\
& \leq \mathbb{E} \left[\sum_{t=t_i}^T I_{\Delta_x < 2\lambda \cdot \max(\epsilon_{k,t}, \epsilon_{l,t})} \cdot I_{n_{x,t} \geq \beta_x \cdot \ln(t)} \right] \\
& \leq \mathbb{E} \left[\sum_{t=t_i}^T I_{\min(n_{k,t}, n_{l,t}) \leq \alpha_x \cdot \ln(t)} \cdot I_{n_{x,t} \geq \beta_x \cdot \ln(t)} \right] \\
& \leq \mathbb{E} \left[\sum_{t=t_i}^T I_{\min(n_{k,t}, n_{l,t}) \leq \alpha_x \cdot \ln(t)} \cdot I_{n_{x,t} \geq \beta_x \cdot \ln(t)} \cdot I_{\bar{c}_{k,t_i} \geq b \geq \bar{c}_{l,t_i}} \right] \\
& \quad + \mathbb{E} \left[\sum_{t=t_i}^T I_{\bar{c}_{k,t_i} < b} \right] + \mathbb{E} \left[\sum_{t=t_i}^T I_{\bar{c}_{l,t_i} > b} \right] \\
& \leq \sum_{t=t_i}^T \mathbb{P}[n_{l,t} \leq \alpha_x \cdot \ln(t) ; n_{x,t} \geq \beta_x \cdot \ln(t) ; \bar{c}_{k,t_i} \geq b \geq \bar{c}_{l,t_i}] \\
& \quad + \sum_{t=t_i}^T \mathbb{P}[n_{k,t} \leq \alpha_x \cdot \ln(t) ; n_{x,t} \geq \beta_x \cdot \ln(t) ; \bar{c}_{k,t_i} \geq b \geq \bar{c}_{l,t_i}] + 2,
\end{aligned}$$

where the last inequality is derived with Lemma 1. Observe that $\frac{\alpha_x}{\beta_x}$ is a constant factor independent of Δ_x . It thus remains to show that if x has been selected at least $\beta_x \ln(t)$ times, then both k and l have been pulled at least a constant fraction of the time with high probability. This is the only time the load balancing algorithm comes into play in the proof of Lemma 14. We study the first term, and we will conclude the study by symmetry. We have:

$$\begin{aligned}
& \mathbb{P}[n_{l,t} \leq \alpha_x \cdot \ln(t) ; n_{x,t} \geq \beta_x \cdot \ln(t) ; \bar{c}_{k,t_i} \geq b \geq \bar{c}_{l,t_i}] \\
& \leq \mathbb{P}[n_{l,t}^x \leq \alpha_x \cdot \ln(t) ; n_{x,t} \geq \beta_x \cdot \ln(t) ; \bar{c}_{k,t_i} \geq b \geq \bar{c}_{l,t_i}] \\
& \leq \sum_{s=\beta_x \cdot \ln(t)}^t \mathbb{P}[n_{l,t}^x \leq \alpha_x \cdot \ln(t) ; n_{x,t} = s ; \bar{c}_{k,t_i} \geq b \geq \bar{c}_{l,t_i}].
\end{aligned} \tag{31}$$

Observe that, by definition of the load balancing algorithm and since $\bar{c}_{k,t_i} \geq b \geq \bar{c}_{l,t_i}$, we are led to pull arm k (resp. arm l) at time t if the budget spent so far when selecting basis x , denoted by b_t^x , is below (resp. above) the “target” of $n_{x,t} \cdot b$. Let us denote by t_1, \dots, t_s the times at which basis x is selected and let us define $(T_n^k)_n$ in $\{1, \dots, s\}$ such as, at times $(t_{T_n^k})_n$, we switch from pulling arm l to pulling arm k , where n identifies the n th switch. We define $(T_n^l)_n$ symmetrically. To simplify the analysis, assume that we start by pulling arm l . Remark that, for any n , we must have:

$$n_{x,t_{T_n^k}} \cdot b \geq b_{t_{T_n^k}}^x \geq n_{x,t_{T_n^k}} \cdot b - b,$$

and:

$$n_{x,t_{T_n^l}} \cdot b + (1 - b) \geq b_{t_{T_n^l}}^x \geq n_{x,t_{T_n^l}} \cdot b,$$

since $c_{l,t}, c_{k,t} \in [0, 1]$. From these two inequalities, we derive:

$$\sum_{i=T_n^k}^{T_n^l-1} c_{k,t_i} = b_{t_{T_n^l}}^x - b_{t_{T_n^k}}^x \leq (T_n^l - T_n^k) \cdot b + 2 \quad \forall n,$$

since $b \in (0, 1)$. If the last switch, n^* , is a $l \rightarrow k$ switch, we get:

$$\sum_{i=T_{n^*}^k}^s c_{k,t_i} < (s - T_{n^*}^k) \cdot b + 2.$$

Summing up these inequalities across all n 's, we obtain:

$$\sum_{i \mid k \text{ is pulled}} c_{k,t_i} < n_{k,t}^x \cdot b + 2 \cdot n_{l,t}^x.$$

We derive:

$$\begin{aligned} & \mathbb{P}[n_{l,t}^x \leq \alpha_x \cdot \ln(t) ; n_{x,t} = s ; \bar{c}_{k,t_i} \geq b \geq \bar{c}_{l,t_i}] \\ & \leq \sum_{z=0}^{\alpha_x \ln(t)} \mathbb{P}[n_{l,t}^x = z ; n_{x,t} = s ; \bar{c}_{k,t_i} \geq b \geq \bar{c}_{l,t_i}] \\ & \leq \sum_{z=0}^{\alpha_x \ln(t)} \mathbb{P}\left[\sum_{i \mid k \text{ is pulled}} c_{k,t_i} < (s - z) \cdot b + 2z ; n_{l,t}^x = z ; n_{x,t} = s\right] \\ & \leq \sum_{z=0}^{\alpha_x \ln(t)} \mathbb{P}\left[\sum_{i \mid k \text{ is pulled}} c_{k,t_i} < (s - z) \cdot \mu_k^c - [(s - z) \cdot \epsilon - 2z] ; n_{l,t}^x = z ; n_{x,t} = s\right] \\ & \leq \sum_{z=0}^{\alpha_x \ln(t)} \exp\left(-2 \frac{((s - z) \cdot \epsilon - 2z)^2}{s - z}\right) \\ & \leq \exp(-2s \cdot \epsilon^2) \cdot \sum_{z=0}^{\alpha_x \ln(t)} \exp(2\epsilon(\epsilon + 4)z) \\ & \leq \exp(-2s \cdot \epsilon^2) \cdot \frac{\exp(2\epsilon(\epsilon + 4) \cdot (\alpha_x \ln(t) + 1))}{\exp(2\epsilon(\epsilon + 4)) - 1}, \end{aligned}$$

where we use Lemma 1. Plugging this last inequality back into (31), we obtain:

$$\begin{aligned} & \mathbb{P}[n_{l,t} \leq \alpha_x \cdot \ln(t) ; n_{x,t} \geq \beta_x \cdot \ln(t) ; \bar{c}_{k,t_i} \geq b \geq \bar{c}_{l,t_i}] \\ & \leq \frac{\exp(-2\epsilon \cdot (\epsilon \cdot \beta_x - (\epsilon + 4) \cdot \alpha_x) \cdot \ln(t))}{(1 - \exp(-2\epsilon^2)) \cdot (1 - \exp(-2\epsilon(\epsilon + 4)))} \\ & \leq \frac{1}{(1 - \exp(-2\epsilon^2)) \cdot (1 - \exp(-8\epsilon))} \cdot \frac{1}{t^2}, \end{aligned}$$

since, by definition of β_x and α_x , we have:

$$\begin{aligned} \frac{1}{\epsilon^2} + \left(\frac{4}{\epsilon} + 1\right) \cdot \alpha_x &\leq \left(\frac{1}{\epsilon^4} + 8\left(\frac{4}{\epsilon} + 1\right) \cdot \lambda^2\right) \cdot \left(\frac{1}{\Delta_x}\right)^2 \\ &\leq \left(\frac{1}{\epsilon^4} + \frac{80}{\epsilon^3}\right) \cdot \left(\frac{1}{\Delta_x}\right)^2 \\ &\leq \beta_x, \end{aligned}$$

where we use the fact that $\epsilon \geq 1$ and $\Delta_x \leq \text{obj}_{x^*} \leq \frac{1}{\epsilon} \sum_{k=1} \mu_k^c \cdot \xi_k^{x^*} \leq \frac{1}{\epsilon}$. We finally conclude that:

$$\begin{aligned} \sum_{t=t_i}^T \mathbb{P}[n_{l,t} \leq \alpha_x \cdot \ln(t) ; n_{x,t} \geq \beta_x \cdot \ln(t) ; \bar{c}_{k,t_i} \geq b \geq \bar{c}_{l,t_i}] \\ \leq \frac{\pi^2}{6 \cdot (1 - \exp(-2\epsilon^2)) \cdot (1 - \exp(-8\epsilon))} \\ \leq \frac{4\pi^2}{6\epsilon^3} \\ \leq \frac{\pi^2}{\epsilon^3}, \end{aligned}$$

since $\exp(-8u), \exp(-2u) \leq 1 - \frac{u}{2}$ for $u \in [0, 1]$. ■

Fact 11

$$\mathbb{E}\left[\sum_{t=t_i}^T I_{\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}} \cdot I_{x \in \mathcal{B}_t}\right] \leq \pi^2 \cdot \left(\frac{1}{3\epsilon^2} + 1\right).$$

Proof Without loss of generality, we can write $x = \{k, l\}$ with $\mu_k^c > b > \mu_l^c$. First observe that:

$$\begin{aligned} \mathbb{E}\left[\sum_{t=t_i}^T I_{\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}} \cdot I_{x \in \mathcal{B}_t}\right] &\leq \mathbb{E}\left[\sum_{t=t_i}^T I_{\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}} \cdot I_{\bar{c}_{k,t_i} \geq b \geq \bar{c}_{l,t_i}}\right] \\ &\quad + \mathbb{E}\left[\sum_{t=t_i}^T I_{\bar{c}_{k,t} < b}\right] + \mathbb{E}\left[\sum_{t=t_i}^T I_{\bar{c}_{l,t} > b}\right] \\ &\leq \mathbb{E}\left[\sum_{t=t_i}^T I_{\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}} \cdot I_{\bar{c}_{k,t} \geq b \geq \bar{c}_{l,t}}\right] + \frac{\pi^2}{3\epsilon^2}, \end{aligned}$$

where we bound the last two terms in the same fashion as in Fact 9. The key observation is that if $\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}$ and $\bar{c}_{k,t_i} \geq b \geq \bar{c}_{l,t_i}$, at least one of the following six events occurs: $\{\bar{r}_{k,t} \geq \mu_k^r + \epsilon_{k,t}\}$, $\{\bar{r}_{l,t} \geq \mu_l^r + \epsilon_{l,t}\}$, $\{\bar{c}_{k,t} \leq \mu_k^c - \epsilon_{k,t}\}$, $\{\bar{c}_{k,t} \geq \mu_k^c + \epsilon_{k,t}\}$, $\{\bar{c}_{l,t} \leq \mu_l^c - \epsilon_{l,t}\}$ or

$\{\bar{c}_{l,t} \geq \mu_l^c + \epsilon_{l,t}\}$. *Otherwise, we have:*

$$\begin{aligned}
 \text{obj}_{x,t} - \text{obj}_x &= \left[\frac{\bar{c}_{k,t} - b}{\bar{c}_{k,t} - \bar{c}_{l,t}} \cdot \bar{r}_{l,t} + \frac{b - \bar{c}_{l,t}}{\bar{c}_{k,t} - \bar{c}_{l,t}} \cdot \bar{r}_{k,t} \right] - \left[\frac{\mu_k^c - b}{\mu_k^c - \mu_l^c} \cdot \mu_l^r + \frac{b - \mu_l^c}{\mu_k^c - \mu_l^c} \cdot \mu_k^r \right] \\
 &< \left[\frac{\bar{c}_{k,t} - b}{\bar{c}_{k,t} - \bar{c}_{l,t}} \cdot (\mu_l^r + \epsilon_{l,t}) + \frac{b - \bar{c}_{l,t}}{\bar{c}_{k,t} - \bar{c}_{l,t}} \cdot (\mu_k^r + \epsilon_{k,t}) \right] \\
 &\quad - \left[\frac{\mu_k^c - b}{\mu_k^c - \mu_l^c} \cdot \mu_l^r + \frac{b - \mu_l^c}{\mu_k^c - \mu_l^c} \cdot \mu_k^r \right] \\
 &= \frac{1}{\lambda} \cdot E_{x,t} + (\mu_k^r - \mu_l^r) \cdot \left[\frac{b - \bar{c}_{l,t}}{\bar{c}_{k,t} - \bar{c}_{l,t}} - \frac{b - \mu_l^c}{\mu_k^c - \mu_l^c} \right] \\
 &= \frac{1}{\lambda} \cdot E_{x,t} + \frac{(\mu_k^r - \mu_l^r)}{(\bar{c}_{k,t} - \bar{c}_{l,t}) \cdot (\mu_k^c - \mu_l^c)} \cdot [(\mu_k^c - b)(\mu_l^c - \bar{c}_{l,t}) + (b - \mu_l^c)(\mu_k^c - \bar{c}_{k,t})] \\
 &< \frac{1}{\lambda} \cdot E_{x,t} + \frac{|\mu_k^r - \mu_l^r|}{(\bar{c}_{k,t} - \bar{c}_{l,t}) \cdot (\mu_k^c - \mu_l^c)} \cdot [(\bar{c}_{k,t} - b) \cdot \epsilon_{l,t} + (b - \bar{c}_{l,t}) \cdot \epsilon_{k,t}] \\
 &= \frac{1}{\lambda} \cdot E_{x,t} + \frac{1}{\epsilon \cdot \lambda} \cdot E_{x,t} \\
 &= E_{x,t},
 \end{aligned}$$

a contradiction. The second inequality is derived from the observation that $(\mu_k^c - b)(\mu_l^c - \bar{c}_{l,t}) + (b - \mu_l^c)(\mu_k^c - \bar{c}_{k,t})$ is a linear function of (μ_k^c, μ_l^c) (since the cross term $\mu_k^c \cdot \mu_l^c$ cancels out) so that the minimum and the maximum of this expression over the polyhedron $[\bar{c}_{k,t} - \epsilon_{k,t}, \bar{c}_{k,t} + \epsilon_{k,t}] \times [\bar{c}_{l,t} - \epsilon_{l,t}, \bar{c}_{l,t} + \epsilon_{l,t}]$ are attained at an extreme point. We obtain:

$$\begin{aligned}
 \mathbb{E} \left[\sum_{t=t_i}^T I_{\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}} \cdot I_{x \in \mathcal{B}_t} \right] &\leq \sum_{t=1}^{\infty} \mathbb{P}[\bar{r}_{k,t} \geq \mu_k^r + \epsilon_{k,t}] + \mathbb{P}[\bar{r}_{l,t} \geq \mu_l^r + \epsilon_{l,t}] \\
 &\quad + \sum_{t=1}^{\infty} \mathbb{P}[\bar{c}_{l,t} \geq \mu_l^c + \epsilon_{l,t}] + \mathbb{P}[\bar{c}_{k,t} \geq \mu_k^c + \epsilon_{k,t}] \\
 &\quad + \sum_{t=1}^{\infty} \mathbb{P}[\bar{c}_{k,t} \leq \mu_k^c - \epsilon_{k,t}] + \mathbb{P}[\bar{c}_{l,t} \leq \mu_l^c - \epsilon_{l,t}] \\
 &\leq \pi^2,
 \end{aligned}$$

using the same argument as in Fact 3. ■

Fact 12

$$\mathbb{E} \left[\sum_{t=t_i}^T I_{\text{obj}_{x^*,t} \leq \text{obj}_{x^*} - E_{x^*,t}} \cdot I_{x^* \in \mathcal{B}_t} \right] \leq \pi^2 \cdot \left(\frac{1}{3\epsilon^2} + 1 \right).$$

We leave out the proof since it is almost identical to the proof of Fact 11.

Appendix N. Proof of Lemma 15

Without loss of generality, we can write $x = \{k, l\}$ with $\mu_k^c > b > \mu_l^c$ and we denote by $(t_n)_{n \in \mathbb{B}}$ the random times at which basis x is selected (these random variables are finite almost surely). Consider

$t \geq t_i$ and $u \geq 1$. Observe that:

$$\begin{aligned} \mathbb{P}[|b_t^x - n_{x,t} \cdot b| \geq u] &\leq \mathbb{P}[|b_t^x - n_{x,t} \cdot b| \geq u ; \bar{c}_{k,t_i} \geq b \geq \bar{c}_{l,t_i}] \\ &\quad + \mathbb{P}[\bar{c}_{k,t_i} \leq b] + \mathbb{P}[\bar{c}_{l,t_i} \geq b] \\ &\leq \mathbb{P}[|b_t^x - n_{x,t} \cdot b| \geq u ; \bar{c}_{k,t_i} \geq b \geq \bar{c}_{l,t_i}] + \frac{2}{T^2}. \end{aligned}$$

using (9) and Lemma 1. Note that, by definition of the load balancing algorithm and since $\bar{c}_{k,t_i} \geq b \geq \bar{c}_{l,t_i}$, we are led to pull arm k (resp. arm l) at time t if the budget spent so far when selecting basis x is below (resp. above) the “target” of $n_{x,t} \cdot b$. Hence, if $b_t^x - n_{x,t} \cdot b \geq u$, we must have been pulling arm l for the last $s \geq \lfloor u \rfloor$ rounds where basis x was selected (because the amounts of resource consumed at each round are almost surely bounded by 1) and we must have:

$$\sum_{n=n_{x,t}-s+1}^{n_{x,t}} c_{l,t_n} \geq s \cdot b + u - 1.$$

Otherwise, since we switched from pulling arm k to pulling arm l at time $t_{n_{x,t}-s+1}$, we have:

$$b_{t_{n_{x,t}-s+1}}^x \leq (n_{x,t} - s) \cdot b + 1,$$

which in combination with:

$$\sum_{n=n_{x,t}-s+1}^{n_{x,t}} c_{l,t_n} < s \cdot b + u - 1$$

yields:

$$b_t^x < n_{x,t} \cdot b + u,$$

a contradiction. Hence, if $u \geq 1$:

$$\begin{aligned} \mathbb{P}[b_t^x - n_{x,t} \cdot b \geq u ; \bar{c}_{k,t_i} \geq b \geq \bar{c}_{l,t_i}] &\leq \sum_{s=\lfloor u \rfloor}^t \mathbb{P}\left[\sum_{n=n_{x,t}-s+1}^{n_{x,t}} c_{l,t_n} \geq s \cdot b + u - 1\right] \\ &= \sum_{s=\lfloor u \rfloor}^t \mathbb{P}\left[\sum_{n=n_{x,t}-s+1}^{n_{x,t}} c_{l,t_n} \geq s \cdot \mu_l^c + (s \cdot (b - \mu_l^c) + u - 1)\right] \\ &\leq \sum_{s=\lfloor u \rfloor}^t \mathbb{P}\left[\sum_{n=n_{x,t}-s+1}^{n_{x,t}} c_{l,t_n} \geq s \cdot \mu_l^c + s \cdot \epsilon\right] \\ &\leq \sum_{s=\lfloor u \rfloor}^t \exp(-2\epsilon^2 \cdot s) \\ &\leq \frac{\exp(-2\epsilon^2 \cdot \lfloor u \rfloor)}{1 - \exp(-2\epsilon^2)} \\ &\leq \frac{2}{\epsilon^2} \cdot \exp(-\epsilon^2 \cdot u), \end{aligned}$$

where we use Lemma 1 for the third inequality and the fact that $\exp(-2v) \leq 1 - \frac{v}{2}$ for $v \in [0, 1]$ for the last inequality. With a similar argument, we conclude:

$$\mathbb{P}[|b_t^x - n_{x,t} \cdot b| \geq u ; \bar{c}_{k,t_i} \geq b \geq \bar{c}_{l,t_i}] \leq \frac{4}{\epsilon^2} \cdot \exp(-\epsilon^2 \cdot u).$$

This last result enables us to show that:

$$\begin{aligned}
 |\mathbb{E}[b_T^x] - \mathbb{E}[n_{x,T}] \cdot b| &\leq \mathbb{E}[|b_T^x - n_{x,T} \cdot b|] \\
 &= \int_0^T \mathbb{P}[|b_T^x - n_{x,T} \cdot b| \geq u] du \\
 &\leq \frac{4}{\epsilon^2} \cdot \int_0^T \exp(-\epsilon^2 \cdot u) du + 1 + \frac{2}{T} \\
 &\leq \frac{4}{\epsilon^4} + 3 \\
 &\leq \frac{8}{\epsilon^4}.
 \end{aligned}$$

We get:

$$\mathbb{E}[n_{k,T}^x] \cdot \mu_k^c + \mathbb{E}[n_{l,T}^x] \cdot \mu_l^c = \mathbb{E}[b_T^x] \geq \mathbb{E}[n_{x,T}] \cdot b - \frac{8}{\epsilon^4},$$

which, in combination with $\mathbb{E}[n_{k,T}^x] + \mathbb{E}[n_{l,T}^x] = \mathbb{E}[n_{x,T}]$, shows that:

$$\begin{aligned}
 \mathbb{E}[n_{k,T}^x] &\geq \left(\frac{b - \mu_l^c}{\mu_k^c - \mu_l^c} \right) \cdot \mathbb{E}[n_{x,T}] - \frac{8}{\epsilon^4 \cdot (\mu_k^c - \mu_l^c)} \\
 &\geq \xi_k^x \cdot \mathbb{E}[n_{x,T}] - \frac{4}{\epsilon^5}.
 \end{aligned}$$

Symmetrically, we get:

$$\mathbb{E}[n_{l,T}^x] \geq \xi_l^x \cdot \mathbb{E}[n_{x,T}] - \frac{4}{\epsilon^5}.$$

Appendix O. Proof of Proposition 16

The proof starts by refining (4):

$$\begin{aligned}
 R_{B,T} &\leq T \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \mathbb{E}\left[\sum_{t=1}^{\tau^*} r_{a_t,t}\right] + O(1) \\
 &= T \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \mathbb{E}\left[\sum_{t=1}^T r_{a_t,t}\right] + \mathbb{E}\left[\sum_{t=\tau^*+1}^T r_{a_t,t}\right] + O(1) \\
 &\leq T \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \mathbb{E}\left[\sum_{t=1}^T r_{a_t,t}\right] + \mathbb{E}\left[\sum_{t=\tau^*+1}^T 1\right] + O(1) \\
 &\leq T \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \mathbb{E}\left[\sum_{t=1}^T r_{a_t,t}\right] + \frac{1}{\epsilon} \cdot \mathbb{E}\left[\sum_{t=\tau^*+1}^T c_{a_t,t}\right] + O(1) \\
 &\leq T \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \mathbb{E}\left[\sum_{t=1}^T r_{a_t,t}\right] + \frac{1}{\epsilon} \cdot \mathbb{E}\left[\left(\sum_{t=1}^T c_{a_t,t} - B\right)_+\right] + O(1).
 \end{aligned}$$

The second inequality holds because the rewards are no greater than 1. The third inequality is a consequence of Assumption 2:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=\tau^*+1}^T c_{a_t,t}\right] &= \mathbb{E}\left[\sum_{t=1}^T c_{a_t,t}\right] - \mathbb{E}\left[\sum_{t=1}^{\tau^*} c_{a_t,t}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}[c_{a_t,t} \mid \mathcal{F}_{t-1}]\right] - \mathbb{E}\left[\sum_{t=1}^{\infty} I_{\tau^* \geq t} \cdot \mathbb{E}[c_{a_t,t} \mid \mathcal{F}_{t-1}]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \mu_{a_t}^c\right] - \mathbb{E}\left[\sum_{t=1}^{\infty} I_{\tau^* \geq t} \cdot \mu_{a_t}^c\right] \\
&= \mathbb{E}\left[\sum_{t=\tau^*+1}^T \mu_{a_t}^c\right] \\
&\geq \mathbb{E}\left[\sum_{t=\tau^*+1}^T \epsilon\right],
\end{aligned}$$

since τ^* is a stopping time. To derive the fourth inequality, observe that if $\tau^* = T + 1$, we have:

$$\sum_{t=\tau^*+1}^T c_{a_t,t} = 0 \leq \left(\sum_{t=1}^T c_{a_t,t} - B\right)_+,$$

while if $\tau^* < T + 1$ we have run out of resources before the end of the game, i.e. $\sum_{t=1}^{\tau^*} c_{a_t,t} \geq B$, which implies:

$$\begin{aligned}
\sum_{t=\tau^*+1}^T c_{a_t,t} &\leq \sum_{t=\tau^*+1}^T c_{a_t,t} + \sum_{t=1}^{\tau^*} c_{a_t,t} - B \\
&\leq \left(\sum_{t=1}^T c_{a_t,t} - B\right)_+.
\end{aligned}$$

Now observe that:

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{t=1}^T c_{a_t,t} - B\right)_+\right] &\leq \mathbb{E}\left[\left(\sum_{t=t_i}^T c_{a_t,t} - (T - t_i + 1) \cdot b\right)_+\right] + t_i \\
&\leq \mathbb{E}\left[\left(\sum_x \{b_T^x - n_{x,T} \cdot b\}\right)_+\right] + \frac{K}{\epsilon^2} \ln(T) + O(1) \\
&\leq \sum_{x \in \mathcal{B}} \mathbb{E}[|b_T^x - n_{x,T} \cdot b|] + \sum_{x \notin \mathcal{B}} \mathbb{E}[n_{x,T}] + \frac{K}{\epsilon^2} \ln(T) + O(1) \\
&\leq \sum_{x \in \mathcal{B}} \int_0^T \mathbb{P}[|b_T^x - n_{x,T} \cdot b| \geq u] du + \frac{K}{\epsilon^2} \ln(T) + O(1) \\
&\leq \frac{4}{\epsilon^2} \cdot \sum_{x \in \mathcal{B}} \int_0^T \exp(-\epsilon^2 \cdot u) du + 1 + \frac{2}{T} + \frac{K}{\epsilon^2} \ln(T) + O(1) \\
&= \frac{4}{\epsilon^4} \cdot |\mathcal{B}| + \frac{K}{\epsilon^2} \ln(T) + O(1),
\end{aligned}$$

where we use the fact that $c_{k,t} \leq 1$ at any time t and for all arms k for the third inequality, Lemma 13 for the fourth inequality, and Lemma 15 for the fifth inequality. Plugging this inequality back into the regret bound yields:

$$\begin{aligned}
 R_{B,T} &\leq T \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \mathbb{E}[\sum_{t=1}^T r_{a_t,t}] + \frac{K}{\epsilon^3} \cdot \ln(T) + O(1) \\
 &\leq T \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \mathbb{E}[\sum_{t=t_i}^T r_{a_t,t}] + \frac{K}{\epsilon^3} \cdot \ln(T) + O(1) \\
 &\leq T \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \sum_{x \in \mathcal{B}} \sum_{k=1}^K \mu_k^r \cdot \mathbb{E}[n_{k,T}^x] + \frac{K}{\epsilon^3} \cdot \ln(T) + O(1) \\
 &\leq T \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \sum_{x \in \mathcal{B}} (\sum_{k=1}^K \mu_k^r \cdot \xi_k^x) \cdot \mathbb{E}[n_{x,T}] + \frac{K}{\epsilon^3} \cdot \ln(T) + O(1) \\
 &\leq \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} \cdot (T - \sum_{x \in \mathcal{B} \mid \Delta_x=0} \mathbb{E}[n_{x,T}]) - \sum_{x \in \mathcal{B} \mid \Delta_x>0} (\sum_{k=1}^K \mu_k^r \cdot \xi_k^x) \cdot \mathbb{E}[n_{x,T}] \\
 &\quad + \frac{K}{\epsilon^3} \cdot \ln(T) + O(1) \\
 &= \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} \cdot (t_i + \sum_{x \in \mathcal{B} \mid \Delta_x>0} \mathbb{E}[n_{x,T}] + \sum_{x \notin \mathcal{B}} \mathbb{E}[n_{x,T}]) - \sum_{x \in \mathcal{B} \mid \Delta_x>0} (\sum_{k=1}^K \mu_k^r \cdot \xi_k^x) \cdot \mathbb{E}[n_{x,T}] \\
 &\quad + \frac{K}{\epsilon^3} \cdot \ln(T) + O(1) \\
 &\leq \sum_{x \in \mathcal{B} \mid \Delta_x>0} \Delta_x \cdot \mathbb{E}[n_{x,T}] + \frac{2K}{\epsilon^3} \cdot \ln(T) + O(1) \tag{32} \\
 &\leq \left(\frac{4}{\epsilon}\right)^4 \cdot \left(\sum_{x \in \mathcal{B} \mid \Delta_x>0} \frac{1}{\Delta_x}\right) \cdot \ln(T) + \frac{2K}{\epsilon^3} \cdot \ln(T) + O(1),
 \end{aligned}$$

where we use Lemma 15 for the fourth inequality, Lemma 13 for the seventh inequality, and Lemma 14 for the last inequality.

Appendix P. Proof of Proposition 17

Along the same lines as for the case of a single limited resource, we start from inequality (32) derived in the proof of Proposition 16 and apply Lemma 14 only if Δ_x is big enough taking into account the fact that:

$$\sum_{x \in \mathcal{B}} \mathbb{E}[n_{x,T}] \leq T.$$

Specifically, we have:

$$\begin{aligned}
R_{B,T} &\leq \sup_{\substack{(n_x)_{x \in \mathcal{B}} \geq 0 \\ \sum_{x \in \mathcal{B}} n_x \leq T}} \left\{ \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \min(\Delta_x \cdot n_x, (\frac{4}{\epsilon})^4 \cdot \frac{\ln(T)}{\Delta_x} + \frac{5\pi^2}{\epsilon^3} \cdot \Delta_x) \right\} + O(1) \\
&\leq \sup_{\substack{(n_x)_{x \in \mathcal{B}} \geq 0 \\ \sum_{x \in \mathcal{B}} n_x \leq T}} \left\{ \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \min(\Delta_x \cdot n_x, (\frac{4}{\epsilon})^4 \cdot \frac{\ln(T)}{\Delta_x}) \right\} + O(1) \\
&= \sup_{\substack{(n_x)_{x \in \mathcal{B}} \geq 0 \\ \sum_{x \in \mathcal{B}} n_x \leq T}} \left\{ \sum_{x \in \mathcal{B}} \sqrt{(\frac{4}{\epsilon})^4 \cdot \ln(T) \cdot n_x} \right\} + O(1) \\
&\leq (\frac{4}{\epsilon})^2 \cdot \sqrt{\ln(T)} \cdot \sup_{\substack{(n_x)_{x \in \mathcal{B}} \geq 0 \\ \sum_{x \in \mathcal{B}} n_x \leq T}} \left\{ \sum_{x \in \mathcal{B}} \sqrt{n_x} \right\} + O(1) \\
&\leq (\frac{4}{\epsilon})^2 \cdot \sqrt{|\mathcal{B}| \cdot T \cdot \ln(T)},
\end{aligned}$$

where we maximize over each $\Delta_x \geq 0$ to derive the third inequality and we use Cauchy-Schwartz for the last inequality.

Appendix Q. Proof of Lemma 18

The proof follows the same steps as for Lemma 10. We use the shorthand $\beta_k = 8 \cdot \frac{r_{1,\dots,C}}{\epsilon^2 \cdot (\Delta_k)^2}$. First observe that:

$$\mathbb{E}[n_{k,\tau^*}] \leq 2\beta_k \cdot \mathbb{E}[\ln(\tau^*)] + \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{a_t=k} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}\right],$$

and we can focus on bounding the second term, which can be broken down as follows:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{a_t=k} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}\right] &\leq \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x_t,t} + E_{x_t,t} \geq \text{obj}_{x^*,t} + E_{x^*,t}} \cdot I_{a_t=k} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x_t,t} \geq \text{obj}_{x_t} + E_{x_t,t}}\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x^*,t} \leq \text{obj}_{x^*} - E_{x^*,t}}\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x^*} < \text{obj}_{x_t} + 2E_{x_t,t}} \cdot I_{a_t=k} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}\right].
\end{aligned}$$

We study each term separately, just like in Lemma 10.

Fact 13

$$\mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x_t,t} \geq \text{obj}_{x_t} + E_{x_t,t}}\right] \leq K \cdot \frac{\pi^2}{6}.$$

Proof If $\text{obj}_{x_t,t} \geq \text{obj}_{x_t} + E_{x_t,t}$, there must exist $l \in \text{supp}(x_t)$ such that $\bar{r}_{l,t} \geq \mu_l^r + \epsilon_{l,t}$, otherwise:

$$\begin{aligned} \text{obj}_{x_t,t} - \text{obj}_{x_t} &= \sum_{l \in \text{supp}(x_t)} (\bar{r}_{l,t} - \mu_l^r) \cdot \xi_l^{x_t} \\ &< \sum_{l \in \text{supp}(x_t)} \epsilon_{l,t} \cdot \xi_l^{x_t} \\ &< E_{x_t,t}. \end{aligned}$$

We obtain:

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x_t,t} \geq \text{obj}_{x_t} + E_{x_t,t}}\right] &\leq \mathbb{E}\left[\sum_{t=1}^{\tau^*} \sum_{l=1}^K I_{\bar{r}_{l,t} \geq \mu_l^r + \epsilon_{l,t}}\right] \\ &\leq \sum_{l=1}^K \sum_{t=1}^{\infty} \mathbb{P}[\bar{r}_{l,t} \geq \mu_l^r + \epsilon_{l,t}] \\ &\leq K \cdot \frac{\pi^2}{6}, \end{aligned}$$

where the last inequality is derived along the same lines as in the proof of Fact 3. ■

Similarly, we can show that:

$$\mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x^*,t} \leq \text{obj}_{x^*} - E_{x^*,t}}\right] \leq K \cdot \frac{\pi^2}{6}.$$

We move on to study the last term.

Fact 14

$$\mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x^*} < \text{obj}_{x_t} + 2E_{x_t,t}} \cdot I_{a_t=k} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}\right] = 0.$$

Proof If $a_t = k$, it must be that we have selected a suboptimal basis at time t since $k \notin x^*$. If moreover $\text{obj}_{x^*} < \text{obj}_{x_t} + 2E_{x_t,t}$, we have:

$$\begin{aligned} \frac{\Delta_k}{2} &\leq \frac{\Delta_{x_t}}{2} \\ &< \sum_{l \in \text{supp}(x_t)} \xi_l^{x_t} \cdot \sqrt{\frac{2 \ln(t)}{n_{l,t}}} \\ &\leq \sum_{l \in \text{supp}(x)} \sqrt{\frac{2 \xi_l^{x_t} \cdot \xi_k^{x_t} \ln(t)}{n_{k,t}}} \end{aligned}$$

where we use the fact that, by definition of the load balancing algorithm and since $a_t = k$:

$$n_{l,t} \geq \frac{\xi_l^{x_t}}{\xi_k^{x_t}} n_{k,t}, \tag{33}$$

for all arms $l \in \text{supp}(x_t)$. We get:

$$\begin{aligned}
n_{k,t} &< \frac{8}{(\Delta_k)^2} \cdot \xi_k^{x_t} \cdot \left(\sum_{l \in \text{supp}(x_t)} \sqrt{\xi_l^{x_t}} \right)^2 \cdot \ln(t) \\
&\leq \frac{8}{(\Delta_k)^2} \cdot \xi_k^{x_t} \cdot r_{1,\dots,C} \cdot \sum_{l \in \text{supp}(x_t)} \xi_l^{x_t} \cdot \ln(t) \\
&\leq \frac{8}{(\Delta_k)^2} \cdot r_{1,\dots,C} \cdot \left(\sum_{l \in \text{supp}(x_t)} \xi_l^{x_t} \right)^2 \cdot \ln(t),
\end{aligned}$$

using the Cauchy–Schwarz inequality and the fact that a basis involves at most $r_{1,\dots,C}$ arms. Now observe that:

$$\begin{aligned}
\sum_{l \in \text{supp}(x_t)} \xi_l^{x_t} &\leq \min_i \sum_{l \in \text{supp}(x_t)} \frac{c_l(i)}{\epsilon} \cdot \xi_l^{x_t} \\
&\leq \frac{\min_i b(i)}{\epsilon} \\
&\leq \frac{1}{\epsilon},
\end{aligned}$$

as x_t is feasible. We obtain:

$$\begin{aligned}
n_{k,t} &< 8 \cdot \frac{r_{1,\dots,C}}{\epsilon^2 \cdot (\Delta_k)^2} \cdot \ln(t) \\
&= \beta_k \cdot \ln(t).
\end{aligned}$$

■

Appendix R. Proof of Lemma 19

The proof is almost the same as for Lemma 18. We use the shorthand notations $\beta_k = 8 \cdot \frac{r_{1,\dots,C}}{\epsilon^2 \cdot (\Delta_k)^2}$ and $n_{k,t}^{\neq x^*} = \sum_{x \in \mathcal{B} \mid k \in x, x \neq x^*} n_{k,t}^x$. We have:

$$\mathbb{E}[n_{k,\tau^*}^{\neq x^*}] \leq 2\beta_k \cdot \mathbb{E}[\ln(\tau^*)] + \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{x_t \neq x^*} \cdot I_{a_t=k} \cdot I_{n_{k,t}^{\neq x^*} \geq \beta_k \ln(t)}\right],$$

and we can focus on the second term, which we can break down as follows:

$$\begin{aligned}
 & \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{x_t \neq x^*} \cdot I_{a_t=k} \cdot I_{n_{k,t}^{\neq x^*} \geq \beta_k \ln(t)}\right] \\
 & \leq \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x_t,t} + E_{x_t,t} \geq \text{obj}_{x^*,t} + E_{x^*,t}} \cdot I_{x_t \neq x^*} \cdot I_{a_t=k} \cdot I_{n_{k,t}^{\neq x^*} \geq \beta_k \ln(t)}\right] \\
 & \leq \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x_t,t} \geq \text{obj}_{x_t} + E_{x_t,t}}\right] \\
 & \quad + \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x^*,t} \leq \text{obj}_{x^*} - E_{x^*,t}}\right] \\
 & \quad + \mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x^*} < \text{obj}_{x_t} + 2E_{x_t,t}} \cdot I_{x_t \neq x^*} \cdot I_{a_t=k} \cdot I_{n_{k,t}^{\neq x^*} \geq \beta_k \ln(t)}\right].
 \end{aligned}$$

Using the exact same proof as in Lemma 18, we can show that:

$$\mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x_t,t} \geq \text{obj}_{x_t} + E_{x_t,t}}\right] \leq K \cdot \frac{\pi^2}{6}$$

and

$$\mathbb{E}\left[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x^*,t} \leq \text{obj}_{x^*} - E_{x^*,t}}\right] \leq K \cdot \frac{\pi^2}{6}.$$

Moreover, just like in Lemma 18, the last term is still equal to zero and the proof is along the same lines except for inequality (33), which now becomes:

$$n_{l,t} \geq \frac{\xi_l^{x_t}}{\xi_k^{x_t}} \cdot n_{k,t} \geq \frac{\xi_l^{x_t}}{\xi_k^{x_t}} \cdot n_{k,t}^{\neq x^*},$$

for all arms $l \in \text{supp}(x_t)$, and the rest follows through.

Appendix S. Proof of Lemma 20

We first show the second inequality by induction on t . The base case is straightforward, suppose that the inequality holds at time $t - 1$. There are three cases:

- arm k is not pulled at time $t - 1$, in which case the left-hand side of the inequality remains unchanged while the right-hand side can only increase hence the inequality holds at time t ,
- arm k is pulled at time $t - 1$ after selecting $x_{t-1} \neq x^*$, in which case both sides of the inequality increase by one and the inequality holds at time t ,

- arm k is pulled at time $t - 1$ after selecting $x_{t-1} = x^*$. First observe that there must exist $l \in \text{supp}(x^*)$ such that $n_{l,t-1} \leq (t - 1) \cdot \frac{\xi_l^{x^*}}{\sum_{r=1}^K \xi_r^{x^*}}$. Suppose otherwise, we have:

$$\begin{aligned}
t - 1 &= \sum_{l=1}^K n_{l,t} \\
&\geq \sum_{l \in \text{supp}(x^*)} n_{l,t} \\
&> \sum_{l \in \text{supp}(x^*)} (t - 1) \cdot \frac{\xi_l^{x^*}}{\sum_{r=1}^K \xi_r^{x^*}} \\
&= t - 1,
\end{aligned}$$

a contradiction. Suppose now by contradiction that inequality (12) no longer holds at time t , we have:

$$\begin{aligned}
n_{k,t-1} &= n_{k,t} - 1 \\
&> n_{x^*,t} \cdot \frac{\xi_k^{x^*}}{\sum_{l=1}^K \xi_l^{x^*}} + \sum_{x \in \mathcal{B}, x \neq x^*} n_{x,t} \\
&\geq (n_{x^*,t} + \sum_{x \in \mathcal{B}, x \neq x^*} n_{x,t}) \cdot \frac{\xi_k^{x^*}}{\sum_{l=1}^K \xi_l^{x^*}} \\
&= (t - 1) \cdot \frac{\xi_k^{x^*}}{\sum_{l=1}^K \xi_l^{x^*}},
\end{aligned}$$

which implies, using the preliminary remark above, that $\frac{n_{k,t-1}}{\xi_k^{x^*}} > \min_{l \in \text{supp}(x^*)} \frac{n_{l,t-1}}{\xi_l^{x^*}}$, a contradiction given the choice of the load balancing algorithm.

We conclude that inequality (12) holds for all times t and arms $k \in \text{supp}(x^*)$. We also derive inequality (11) as a byproduct, since, at any time t and for any arm $k \in \text{supp}(x^*)$:

$$\begin{aligned}
n_{k,t} &\geq n_{x^*,t} - \sum_{l \in \text{supp}(x^*), l \neq k} n_{l,t} \\
&\geq n_{x^*,t} \cdot \left(1 - \frac{\sum_{l \in \text{supp}(x^*), l \neq k} \xi_l^{x^*}}{\sum_{l=1}^K \xi_l^{x^*}}\right) - r_{1,\dots,C} \cdot \left(\sum_{x \in \mathcal{B}, x \neq x^*} n_{x,t} + 1\right) \\
&= n_{x^*,t} \cdot \frac{\xi_k^{x^*}}{\sum_{l=1}^K \xi_l^{x^*}} - r_{1,\dots,C} \cdot \left(\sum_{x \in \mathcal{B}, x \neq x^*} n_{x,t} + 1\right),
\end{aligned}$$

as basis x^* involves at most $r_{1,\dots,C}$ arms.

Appendix T. Proof of Proposition 21

The proof proceeds along the same lines as for Proposition 11. We start by refining (4):

$$\begin{aligned}
 R_{B(1), \dots, B(C)} &\leq B \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \sum_{k=1}^K \mu_k^r \cdot \mathbb{E}[n_{k, \tau^*}] + O(1) \\
 &\leq B \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \sum_{k \in \text{supp}(x^*)} \mu_k^r \cdot \mathbb{E}[n_{k, \tau^*}] + O(1) \\
 &\leq (B - \frac{\mathbb{E}[n_{x^*, \tau^*}]}{\sum_{k \in \text{supp}(x^*)} \xi_k^{x^*}}) \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} + (r_{1, \dots, C})^2 \cdot \sum_{x \in \mathcal{B}, x \neq x^*} \mathbb{E}[n_{x, t}] + O(1),
 \end{aligned}$$

where we use (11) to derive the third inequality. Now observe that, by definition, at least one resource is exhausted at time τ^* . Hence, there exists $i \in \{1, \dots, C\}$ such that the following holds almost surely:

$$\begin{aligned}
 B(i) &\leq \sum_{k=1}^K c_k(i) \cdot n_{k, \tau^*} \\
 &\leq \sum_{k \notin \text{supp}(x^*)} n_{k, \tau^*} + \sum_{k \in \text{supp}(x^*)} c_k(i) \cdot n_{k, \tau^*} \\
 &\leq \sum_{x \in \mathcal{B}, x \neq x^*} n_{x, \tau^*} + \sum_{k \in \text{supp}(x^*)} c_k(i) \cdot n_{k, \tau^*} \\
 &\leq r_{1, \dots, C} \cdot \left(\sum_{x \in \mathcal{B}, x \neq x^*} n_{x, \tau^*} + 1 \right) + n_{x^*, \tau^*} \cdot \sum_{k \in \text{supp}(x^*)} c_k(i) \cdot \frac{\xi_k^{x^*}}{\sum_{l=1}^K \xi_l^{x^*}} \\
 &\leq r_{1, \dots, C} \cdot \left(\sum_{x \in \mathcal{B}, x \neq x^*} n_{x, \tau^*} + 1 \right) + b(i) \cdot \frac{n_{x^*, \tau^*}}{\sum_{l=1}^K \xi_l^{x^*}},
 \end{aligned}$$

where we use inequality (12) and the fact that basis x^* is feasible for (3). Rearranging yields:

$$\frac{n_{x^*, \tau^*}}{\sum_{l=1}^K \xi_l^{x^*}} \geq B - \frac{r_{1, \dots, C} \cdot (\sum_{x \in \mathcal{B}, x \neq x^*} n_{x, \tau^*} + 1)}{b},$$

almost surely. Plugging this last inequality back into the regret bound, we get:

$$\begin{aligned}
R_{B(1), \dots, B(C)} &\leq r_{1, \dots, C} \cdot \sum_{x \in \mathcal{B}, x \neq x^*} \mathbb{E}[n_{x,t}] \cdot \left(\frac{\sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*}}{b} + r_{1, \dots, C} \right) + O(1) \\
&\leq r_{1, \dots, C} \cdot \sum_{x \in \mathcal{B}, x \neq x^*} \mathbb{E}[n_{x,t}] \cdot \left(\frac{\min_i \sum_{k=1}^K c_k(i) \cdot \xi_k^{x^*}}{\epsilon \cdot b} + r_{1, \dots, C} \right) + O(1) \\
&\leq \left(\frac{r_{1, \dots, C}}{\epsilon \cdot b} + (r_{1, \dots, C})^2 \right) \cdot \sum_{x \in \mathcal{B}, x \neq x^*} \mathbb{E}[n_{x,t}] + O(1) \\
&= \left(\frac{r_{1, \dots, C}}{\epsilon \cdot b} + (r_{1, \dots, C})^2 \right) \cdot \left(\mathbb{E} \left[\sum_{k \in \text{supp}(x)} \sum_{x \in \mathcal{B} \mid k \in x, x \neq x^*} n_{k, \tau^*}^x \right] + \mathbb{E} \left[\sum_{k \notin \text{supp}(x^*)} n_{k,t} \right] \right) + O(1) \\
&\leq \frac{16 \cdot (r_{1, \dots, C})^3}{\epsilon^3 \cdot b} \cdot \left(\sum_{k=1}^K \frac{1}{(\Delta_k)^2} \right) \cdot \mathbb{E}[\ln(\tau^*)] + O(1) \\
&\leq \frac{16 \cdot (r_{1, \dots, C})^3}{\epsilon^3 \cdot b} \cdot \left(\sum_{k=1}^K \frac{1}{(\Delta_k)^2} \right) \cdot \ln\left(\frac{B+1}{\epsilon}\right) + O(1),
\end{aligned}$$

where we use the fact that basis x^* is feasible for (3) for the third inequality, Lemmas 18 and 19 for the fourth inequality, the concavity of the logarithmic function along with Lemma 9 for the last inequality.